UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE FILOSOFÍA Y LETRAS
INSTITUTO DE INVESTIGACIONES FILOSÓFICAS

EL CONCEPTO DE *IMPLEMENTACIÓN* EN LA TEORÍA DE LA CONCIENCIA DE CHALMERS

TESIS

QUE PARA OPTAR POR EL GRADO DE
MAESTRÍA EN FILOSOFÍA DE LA CIENCIA

PRESENTA

ANTONIO GONZÁLEZ GARCÍA

DIRECTOR: DR. FRANCISCO HERNÁNDEZ QUIROZ

ENERO DE 2009

Para:

Araceli, mi familia, mis compañeros y amigos de la UNAM.

Agradezco especialmente al Dr. Francisco Hernández y al Dr. José Luis Díaz por sus enseñanzas y consejos.

Este trabajo se llevó a cabo gracias al apoyo brindado por CONACYT durante el periodo de Septiembre de 2006 a Agosto de 2008 y al Programa de Fomento a la Graduación de la Coordinación de Estudios de Posgrado de la UNAM durante el periodo de Octubre a Diciembre de 2008.

Índice

Introducción

Entre los estudiosos de la filosofía se suele considerar que se dividió el abordaje de los problemas relacionados con la mente en los aspectos científico y filosófico a partir de los textos de Descartes. Con el paso del tiempo y hasta mediados del siglo XX, estos aspectos se alejaron cada vez más uno del otro hasta el punto de considerarse como problemas independientes, a pesar de compartir su objeto principal de estudio: la mente humana. Así, el aspecto científico se ocupaba de problemas relacionados con la fisiología y el comportamiento, mientras que el aspecto filosófico se ocupaba en proveer explicaciones metafísicas y epistemológicas acerca de la naturaleza de la mente. Sin embargo, el desarrollo tanto de la ciencia como de la filosofía provocó que los aspectos anteriores convergieran en ciertos puntos, favoreciendo el estudio de la mente (y como consecuencia el estudio de la conciencia) mediante un enfoque interdisciplinario, provocando un reencuentro entre los abordajes mencionados aparentemente antagónicos. Aunado a esto, la aparición de nuevas disciplinas relacionadas como la teoría de la información y la ciencia computacional, propició la generación de numerosas teorías con expectativas ambiciosas, siendo tarea de algunas de ellas resaltar la importancia del enfoque científico disminuyendo la del enfoque filosófico y viceversa.

La historia de la llamada *ciencia cognitiva* ofrece un ejemplo de lo anterior. Desde mediados del siglo XIX, algunas disciplinas como la neuropsicología y la psicofisiología, han contribuido a establecer una relación entre la psicología y la neurofisiología. Sin embargo, el advenimiento del conductismo y su papel como corriente predominante en la psicología a mediados del siglo XX, establece una temporal separación en el interés de la psicología y la neurofisiología. Lo anterior sucedió como consecuencia de que enfoque el conductista considera que los componentes relevantes de la psicología son sus partes tangibles: el estímulo y la conducta resultante, mientras que los mecanismos neurofisiológicos son considerados como una *caja negra*, donde la entrada es el estímulo y la salida es la conducta

asociada, y no es necesario conocer la estructura de los mecanismos de dicha caja para entender y explicar los mecanismos psicológicos.

Por ello, a pesar de los resultados obtenidos por la neuropsicología y la psicofisiología, los avances en el estudio de la psicología no se encontraban necesariamente ligados al avance en la neurofisiología. Sin embargo, en los años cincuentas se dieron las condiciones adecuadas para que la psicología se aventurara a horizontes distintos al conductismo, reconsiderando y enriqueciendo diversos enfoques filosóficos; todo esto atado fuertemente por las posibilidades que presentaba el uso de la computadora como herramienta experimental, dando origen a una visión multidisciplinaria de los problemas relacionados con la mente conocida como *ciencia cognitiva*.

Con la psicología ocupada una vez más en los conceptos mentales, se buscaron los medios experimentales para definir y mesurar el efecto de un fenómeno mental inaccesible a través de la conducta. Así se inicia el estudio de aspectos no conductuales como la percepción, la memoria, la formación de conceptos, el lenguaje, etcétera. De algunos de esos fenómenos, existían ya referencias de los mecanismo neurofisiologicos involucrados (como las afasias con el lenguaje) gracias a la neurofisiología (Luria, 1978), integrándose de manera automática y estimulando su avance. Por último, la metáfora de la mente como computadora permitía formular hipótesis experimentales e incluso legitimar hasta cierto punto la teoría mediante la simulación en computadoras. Tendido el puente entre esas disciplinas, los fundamentos filosóficos que subyacen a las teorías científicas relacionadas, retoman un papel de gran importancia como sustrato integrador y como una herramienta para evaluar los argumentos aportados por los científicos. Si bien a la fecha difícilmente se puede hablar de una ciencia cognitiva unificada por objetivos comunes y por una comunidad de investigadores organizada, el aspecto interdisciplinario ha producido un fuerte impacto positivo en la mayoría de las disciplinas involucradas (Gardner, 1985).

La razón de la importancia de la teoría que propone David Chalmers para explicar la conciencia radica en su apuesta de equilibrar los

aspectos científico y filosófico inmersos en la ciencia cognitiva: por un lado está una propuesta metafísica que tiene el mérito de tratar de aportar información a las diversas disciplinas científicas involucradas, y por el otro intenta que los datos científicos sirvan también para apoyar o moldear la propuesta metafísica. Este intento convierte a dicha teoría en un posible reconciliador entre los aspectos mencionados, a diferencia de aquellas posturas que niegan cualquier abordaje científico sea posible, y de las que restan importancia a la experiencia subjetiva por la dificultad que implica su tratamiento desde un punto de vista científico. Por otro lado, nuevas disciplinas como la inteligencia artificial se involucran también en los problemas de la conciencia, exigiendo y aportando más precisión y poder explicativo del que se requería anteriormente para postular teorías de la conciencia.

En el artículo llamado *Facing up to the problem of consciousness*, David Chalmers (1995) expone de una manera concisa los postulados filosóficos de su teoría de la conciencia, los cuales desarrolla con mayor profundidad en su libro titulado *The conscious mind* (Chalmers, 1997). En el capítulo 9 de dicho libro, el autor retoma gran parte de lo expuesto en un artículo previo llamado *On implementing a computation* (Chalmers, 1994) para sugerir una aplicación de su teoría de la conciencia en el ámbito de la computación. Allí Chalmers propone que las pretensiones de la llamada inteligencia artificial *fuerte* con respecto a la existencia de un tipo de cómputo suficiente para *ser* una mente, son razonables, y como consecuencia es razonable también la experiencia consciente en las máquinas computadoras. Para afirmar lo anterior, Chalmers utiliza como base los postulados filosóficos centrales de su teoría de la conciencia (la coherencia estructural, la invariancia organizacional y el doble aspecto de la información), además de una manera de concebir el concepto de *implementación*, usándolo como una relación de equivalencia particular entre la estructura causal de un sistema físico o cognitivo y su representación en un cómputo.

El objetivo de este texto será hacer explícitos los presupuestos subyacentes en la aplicación computacional propuesta por Chalmers, especialmente en la concepción de la *inteligencia artificial fuerte* y su compromiso con algunas posiciones representacionalistas de la mente. En el mismo tenor, argumentaré que el uso que Chalmers le da al concepto de *implementación* es atípico y también está fuertemente influenciado por los mismos presupuestos comprometidos con el representacionalismo.

Cabe mencionar que el alcance de esta tesis se limita a extraer los problemas implícitos en la aplicación computacional de la teoría de Chalmers sin argumentar ni a favor ni en contra de los postulados filosóficos que forman el núcleo de la teoría. Entonces, si se cumple el objetivo y se logra demostrar que la aplicación computacional presupone postulados difíciles de aceptar, esto no podría fungir como un argumento en contra de la teoría de la conciencia de Chalmers, sino que tan solo puede aspirar disminuir su posible soporte empírico.

A pesar de lo anterior, al final del texto propondré que gran parte de los presupuestos problemáticos parecen tener origen en un aspecto preciso de la parte filosófica de la teoría: el principio de la invariancia organizacional. Con lo anterior aspiro a esbozar un camino de análisis tanto para la teoría de Chalmers como para cualquier otra teoría de la conciencia donde el concepto de *implementación* sea un aspecto crucial como puente entre el ámbito representacional y el ámbito físico.

Para alcanzar el objetivo mencionado, divido el estudio en cuatro capítulos: el primero funcionará como introducción a las teorías de la conciencia y al contexto general en el que se desarrollan. Así, será necesaria una breve recapitulación del problema de la conciencia y de su desarrollo histórico para mostrar los diferentes puntos de vista y las características compartidas y antagónicas de algunas de las teorías de la conciencia vigentes. En el segundo capítulo ofreceré un breve recorrido histórico de la teoría de la computación, resaltando su relación con la idea del modelado de la mente humana como motivación e incentivo, para después continuar con mi versión de la relación que tiene todo lo anterior con el proyecto de la inteligencia artificial fuerte, así como con la discusión filosófica vigente al respecto. El tercer capítulo será un resumen de los postulados y argumentos filosóficos de la teoría de la conciencia de Chalmers, los cuales, junto a los antecedentes de los dos primeros capítulos, servirán como herramientas para situar a dicha teoría en el panorama general de las teorías de la conciencia. Lo anterior podrá servir también como una herramienta para comprender la relación de la teoría de Chalmers con la inteligencia artificial fuerte. En el último capítulo revisaré la propuesta específica del concepto de *implementación*, para lo cual retomaré los usos tradicionales del término, así como con las motivaciones expuestas en el segundo capítulo y por último ofreceré mi punto de vista al respecto.

1. El problema de la conciencia

Para poder comprender cualquier teoría que pretenda explicar el fenómeno de la conciencia, es necesario conocer un panorama del contexto en donde se genera y desarrolla, además de sus postulados y argumentos. Surge entonces la necesidad de un recuento histórico de las formas en las que se ha abordado el fenómeno de la conciencia, los problemas que han surgido, los que se han solucionado, los que continúan sin solución y los que emergen en el proceso. Es por esto que el objetivo de este primer capítulo es dar un panorama general del problema de la conciencia, desde las primeras inquietudes metafísicas hasta las teorías en auge en la actualidad. De esta manera, en las siguientes páginas presento un resumen de algunos de los momentos más relevantes asociados al problema de la conciencia y cómo se ha concebido tanto en la filosofía como en la ciencia; con la finalidad de comprender, ubicar y evaluar la teoría de la conciencia propuesta por David Chalmers. Lo presentado en éste capítulo será también relevante en las secciones posteriores donde analizaremos si la noción de conciencia resultante de la aplicación de la teoría de Chalmers en un sistema de inteligencia artificial captura lo que históricamente se ha concebido como el fenómeno de la conciencia y, en caso afirmativo, hasta qué grado lo hace.

1.1 Algunos momentos históricos en el problema de la conciencia

La explicación y modelado de lo que entendemos como conciencia representa uno de los más grandes retos tanto para la filosofía como para la ciencia del día de hoy. En palabras de Chalmers (1996, p.1): la conciencia es un misterio. Desde el punto de vista filosófico, el problema de la conciencia desde sus inicios se ha planteando como la manera de establecer y comprender una relación entre dos instancias, reconocidas a lo largo de la historia como espíritu-materia, mente-cuerpo, conciencia-cerebro. Podemos rastrear sus inicios en las tradiciones de los griegos presocráticos y los vedas en la India buscando el establecimiento de la relación entre el espíritu y la

materia (Díaz, 2007). Posteriormente en la tradición europea, el filósofo más representativo que redefine la búsqueda de esta relación es Descartes (1637) cuando se pregunta acerca de la manera en que se relacionan la mente y el cuerpo. Es a partir de ese punto donde podemos reconocer la división franca del estudio filosófico de esta relación en tres tradiciones principales: el dualismo, el idealismo y el materialismo. En la forma más cruda de cada uno de ellos, el dualismo propone que el cuerpo humano y el mundo espiritual se encuentran separados y que interactúan vía ciertos mecanismos, mientras que el idealismo no reconoce que exista materia, sino que todo son formas diversas de las ideas (producto de la mente) y por su parte el materialismo no reconoce la existencia de la mente como sustancia independiente, sino que todo pensamiento o acto mental es una forma de funcionar de la materia.

1.1.1 El tratamiento científico.

Las tradiciones filosóficas anteriormente mencionadas comienzan tratando el problema mente-cuerpo como un estudio principalmente metafísico dada la dificultad de definir un objeto de estudio preciso que pueda tener cabida en la práctica científica de la época. Es hasta la aparición de la psicología y la neurología como actividades científicas independientes (a finales del siglo XIX) cuando se esboza una forma más objetiva de acercarse a este tema, acotando el problema a conceptos más precisos: la relación mente-cerebro. Lo anterior sucede en virtud de que en disciplinas como la neuropsicología, se provee un abordaje empírico de trastornos de la mente asociados con lesiones cerebrales y deja de lado la idea de la época cartesiana de la mente como una entidad ajena al cuerpo. Un ejemplo de lo anterior es el descubrimiento de las funciones especializadas de los hemisferios cerebrales descubiertas gracias a las disfunciones mentales encontradas en los pacientes con el cuerpo calloso cercenado (Sperry, 1977).

Así, la tarea de algunos científicos de fines del siglo XIX como William James era la búsqueda de unidades mínimas en los sistemas mental y cerebral, para de esa manera poder postular teorías acerca de las correlaciones mente-cerebro que eventualmente pudieran derivar en leyes

psicofísicas (James, 1890). Sin embargo, con el advenimiento de las corrientes del conductismo y el psicoanálisis en la psicología de principios del siglo XX y las limitadas herramientas que proveía la neurología en aquel momento, el problema de la conciencia inmerso en la relación mente-cerebro no tuvo cabida como objeto de estudio, relegándose a un cuestionamiento metafísico de poca o nula importancia para la ciencia.

1.1.2 El tratamiento fenomenológico.

En contraste con la tendencia anterior, existe una tradición filosófica también de principios del siglo XX, que plantea la posibilidad de acceder a la conciencia mediante la investigación directa (en primera persona), conocida como *fenomenología*. En general, la fenomenología estudia la experiencia subjetiva y la esencia de lo que experimenta el individuo utilizando el método de la introspección.

Varela, Thompson y Rosch (1992, pp. 15-16) postulan como el principal iniciador de la tradición fenomenológica a Franz Brentano (1874), debido a que a él se le atribuye la noción de *intencionalidad* de los estados mentales, la cual considera que todos los estados mentales necesariamente refieren a un contenido, y tal característica es crucial para la fenomenología. A partir de estas ideas parte Edmund Husserl (1913) para intentar desarrollar un procedimiento para examinar la intencionalidad y la experiencia subjetiva sin apelar al mundo empírico. Posteriormente es Maurice Merleau-Ponty (1942) quien trata de llevar a un punto de convergencia a la fenomenología concebida por Husserl con la ciencia contemporánea, especialmente con la psicología y la neurofisiología. Sin embargo, el proyecto de la fenomenología como ciencia presentó problemas presentes desde su planteamiento original. Estos problemas radican principalmente en que la fenomenología es un abordaje completamente teórico a las cuestiones mentales, y sin una manera adecuada de llevarlo a la práctica, dicha disciplina no tuvo cabida en ciencia de la época. Según Varela, Thompson y Rosch (1992, pp.149), la parte filosófica de la fenomenología trascendió especialmente en la filosofía continental, influenciando fuertemente a autores como Jean Paul Sartre

(1982) y Michael Foucault (1997), y en la parte científica su influencia es limitada, restringiéndose principalmente la psiquiatría y al controvertido[1] psicoanálisis. A pesar de lo anterior, el proyecto original de la fenomenología tal como fue concebido por sus mayores exponentes, es considerado como por algunos como un fracaso (Varela, Thompson y Rosch, 1992, pp. 42-44). Sin embargo, la viva tradición de la fenomenología en la filosofía ha dado lugar a propuestas de reincorporar ciertos aspectos y herramientas fundadas en la fenomenología en diversos ámbitos de la ciencia, como el uso de la introspección como herramienta científica (Díaz, 2007, pp. 323-326; Lutz y Thompson, 2003).

1.1.3 Las nuevas disciplinas y la convergencia .

A partir de mediados del siglo XX, el avance de la neurología y la psiquiatría, aunado a la aparición de nuevas disciplinas como las ciencias cognitivas, la teoría de la información y la computación, dio nueva luz tanto a la forma de abordar el problema de la conciencia, como a las perspectivas que pueden alcanzarse mediante su estudio y su aplicación en la práctica. De esta manera, el aspecto interdisciplinario del problema de la conciencia ha expandido su dominio, exigiendo así que cualquier posición filosófica vaya más allá de una argumentación adecuada y sea capaz de concordar con (o por lo menos no contradecir a) los hechos experimentales, estudiados y explicados por todas las disciplinas científicas relacionadas. Es así como el día de hoy es creciente el interés por el problema de la conciencia y son numerosas las teorías herederas de manera directa o indirecta de alguna de las tradiciones metafísicas mencionadas anteriormente o una variante o mezcla de ellas. Estas teorías presentan un mayor soporte de los datos empíricos de determinada disciplina que de los de alguna otra, dependiendo del enfoque y de la finalidad con la que fueron postuladas. Por ejemplo: es conocimiento común de la neurofisiología que las lesiones en la parte de la corteza cerebral conocida como el área de Wernicke implican una disfunción en la

1 Existe un debate aún abierto acerca del estado del psicoanálisis como elemento perteneciente a la ciencia. Ejemplo de este debate en (Mathers, 1986).

comprensión del lenguaje (Luria, 1978). Este hecho, entre otros, se ha esgrimido como soporte empírico para apoyar la teoría científica de la modularidad de la mente (Fodor, 1983), y ésta a su vez comparte y apoya lo postulado por la teoría filosófica conocida como funcionalismo (Putnam, 1967). El funcionalismo pretende tener un fuerte soporte de diversas disciplinas y esto podría considerarse como una razón importante para enfocar la investigación en el área de la inteligencia artificial guiada por sus postulados (Brooks, 1986) .

1.2 Conciencia psicológica y conciencia fenoménica

Un inconveniente que surge al existir una gama tan amplia de puntos de vista de un problema como el que nos ocupa, es que se tiene la posibilidad de que no sea uno solo, sino varios problemas. Si dichos problemas están englobados en una sentencia muy general, cuando se intentan abordar para su análisis y resolución, esa sentencia general se presenta como un problema ambiguo, que resulta difícil desde su mismo planteamiento y probablemente irresoluble. Así, el punto de partida de la empresa debe ser la definición y acotación de nuestro entendimiento del concepto de conciencia, puesto que tanto en el idioma español como en el inglés y probablemente en otros más, *conciencia* y *consciente* son palabras de uso común tanto en el lenguaje de todos los días como en el científico.

En este tenor, es posible observar que la mayor parte de las veces, el uso del término *conciencia* se refiere a alguno de los siguientes conceptos: vigilia, introspección, informatividad, autoconciencia (o autoreferencia), atención, control voluntario o conocimiento, siendo todos los sustantivos anteriores descripciones funcionales de estados corporales o psicológicos de los individuos. Chalmers (1996) considera a los usos anteriores del término en cuestión como parte de lo que define como *conciencia psicológica* y afirma que los problemas relacionados con ésta eventualmente pueden ser resueltos por la ciencia sin intervención de la filosofía, puesto que no presentan un cuestionamiento metafísico fuerte donde se vuelva indispensable el análisis filosófico. Dado lo anterior, procede a llamarle a

esas áreas del estudio de la conciencia como *el problema fácil*[2]. Así, Chalmers, enfatiza que la parte más compleja del fenómeno de la conciencia es su carácter fenoménico o de la subjetividad de la experiencia, procediendo a llamar a este punto *el problema difícil*, considerándolo como el aspecto medular a resolver en cualquier teoría de la conciencia.

Chalmers menciona que la descripción y caracterización de este último concepto presenta una dificultad intrínseca a su naturaleza subjetiva y por tanto su tratamiento científico es muy delicado (esta es la razón para referirse a él con el apelativo de *problema difícil*). Así, a manera de esclarecer lo que entenderemos como *conciencia fenoménica*, reproduzco una de las descripciones que podría ser considerada como clásica debido a su claridad: la que ofrece Nagel en su artículo *¿Cómo es ser un murciélago?* (Nagel, 1974, p. 46):

> Al margen de cómo varíe la forma [de la experiencia], el hecho que un organismo tenga experiencias conscientes significa básicamente que hay algo que es como es *ser* ese organismo. Puede haber implicaciones sobre la forma de la experiencia; incluso puede haber (aunque lo dudo) implicaciones sobre la conducta del organismo. Pero, fundamentalmente, un organismo tiene estados conscientes si y sólo si hay algo que es cómo es *ser* ese organismo, algo que es cómo es ser *para* ese organismo.

Con el objetivo de facilitar la manera de abordar el problema de la conciencia, retomaré la división del problema de la conciencia propuesta por Chalmers en *conciencia psicológica* y *conciencia fenoménica*. Entonces, cuando el término se refiera a descripciones funcionales de estados

2 La afirmación de que las cuestiones correspondientes a la conciencia psicológica y su esclarecimiento vía la actividad científica (*el problema fácil*) es un debate abierto aún como se puede leer en Lowe (1997) y Hodgson (1997), pero lo relevante es que el lector comprenda la diferencia entre los problemas empíricos, los cuales no son necesariamente fáciles, pero la mayoría son solubles al menos en principio, y el problema filosófico, del cual no estamos seguros si es soluble al menos en principio.

corporales o psicológicos como los enumerados arriba, me referiré a la *conciencia psicológica*, y cuando el término apunte a la experiencia individual tal como es experimentada por el organismo o sistema, hablaremos de *conciencia fenoménica*.

1.3 Objetivos de una teoría de la conciencia

Con las consideraciones anteriores, a continuación propongo los aspectos (tanto psicológicos como fenoménicos) que se deben esperar que abarque una teoría de la conciencia, expresados en cuatro puntos estratégicos. Si bien sabemos que ninguna de las teorías existentes actualmente responde de manera contundente a todas estas exigencias, por lo menos deben ofrecer una posible línea de estudio que eventualmente defina una postura al respecto.

El primer punto se refiere al modelado y caracterización de la conciencia, donde esperamos una respuesta a la pregunta (a) ¿qué es la conciencia? Esto implica que la teoría en cuestión ofrezca una definición concreta de la manera en que se está abordando el tema de la conciencia donde podamos distinguir claramente los componentes, procesos, relaciones y correlaciones que la conforman. Para el segundo punto se espera una explicación acerca de la naturaleza de la conciencia que responda a la pregunta (b) ¿cómo es que existe? Esto significa que, dado el modelo del primer punto, se ofrezca una explicación de la manera en que emerge o existe en los organismos o sistemas a partir de componentes no conscientes. El tercer punto se refiere al estado causal de la conciencia ligado a las preguntas (c) ¿por qué y para qué existe la conciencia?, buscando indagar en la razón de su existencia y si ella implica una diferencia entre la operación de los sistemas que la tienen y los que no, y en caso de ser así, cuál es esa diferencia. El cuarto y último punto corresponde a la explicación de la conciencia fenoménica o al *problema difícil*, tratando de responder a las preguntas (d) ¿por qué y para qué el fenómeno de la conciencia lleva implícita una experiencia subjetiva ligada a cada experimentador individual? De esta manera se busca que toda teoría de la conciencia incluya el carácter

fenoménico de la misma, puesto que si bien es el de más difícil tratamiento desde el punto de vista científico, sin su consideración el problema de la conciencia solo puede abordarse de manera parcial y limitada. Como lo plantea Merleau-Ponty (1945) en su proyecto de fenomenología: la idea es entender al mundo y a la subjetividad como parte inherente del mundo.

1.4 Clasificación de las teorías de la conciencia

Al tenor de estos requerimientos, a continuación propongo una clasificación de las principales teorías filosóficas de la conciencia que actualmente gozan de algún respaldo científico o al menos podrían contar con una actitud favorable por parte de al menos alguna teoría científica dada la compatibilidad de sus postulados. Con ello, el objetivo no es agotar todas las teorías existentes ni profundizar en sus contenidos, sino únicamente proveer un marco de referencia que sea capaz de mostrar la manera de responder cualquier teoría ante las exigencias enumeradas en la sección anterior, lo cual apunta a cumplir con el objetivo general de comprender la teoría de Chalmers antes de analizarla en el capítulo 3.

La delimitación del concepto de conciencia asociada a la pregunta (a) define el criterio más general de esta clasificación, es decir que establece la división entre lo que se aceptará como teoría de la conciencia y por tanto podrá ser clasificado de esta manera. Así, las teorías concernientes a esta clasificación serán aquellas con un enfoque principalmente filosófico que expliquen la conciencia fenoménica (cumpliendo así con el requisito de la pregunta (d)), mientras que las concernientes a la conciencia psicológica (aquellas con una inclinación esencialmente científica) serán consideradas como *soporte empírico indirecto*[3]. De esta forma será la posición asociada a la pregunta (b) el criterio a utilizar para distinguir entre las diferentes clases

3 Con soporte empírico indirecto me refiero a que, para que una teoría científica otorgue respaldo empírico a una teoría filosófica, la primera deberá contar con su propio soporte empírico, el cual será transferido a la segunda. Así, una teoría filosófica A puede casar con un número considerable de teorías científicas, de las cuales todas ellas tienen un soporte empírico pobre, mientras que una teoría filosófica B coincida en sus postulados con una única teoría científica con fuertes fundamentos empíricos. Así, la teoría B tendrá un mayor soporte empírico indirecto que la teoría A.

de teorías, dejando lo concerniente a la pregunta (c) como parte de la tarea de las teorías científicas asociadas a la teoría filosófica en cuestión.

1.4.1 Clasificación según su tradición filosófica

Como punto de partida es necesario recordar las principales tradiciones de la filosofía de la mente y del problema de la conciencia mencionadas anteriormente: el idealismo, el dualismo y el materialismo. Así, en la búsqueda de representantes vigentes de estas tradiciones, se puede considerar que el idealismo, aunque ha sido apoyado científicamente (por ejemplo por interpretaciones extremas de la física cuántica (Gödel, 1949)), en general es difícil que los científicos lo consideren como comprobable o refutable por su naturaleza anti-empírica, y no hay teorías relevantes vigentes que puedan representarlo.

A primera vista, el dualismo parecería destinado a la exclusión del abordaje científico por la misma razón que el idealismo, sin embargo, una modificación en la concepción cartesiana le da cabida en el ámbito de la ciencia. Dicha modificación consiste en dejar de lado la idea de Descartes (1637) de que el cuerpo y la mente están constituidos por *sustancias independientes,* pues dicho postulado acarrea los mismos problemas metafísicos del idealismo. En su lugar, este nuevo tipo de dualismo pretende llevar al terreno de la discusión la naturaleza predicativa y ontológica de las *propiedades* de los mecanismos físico y mental (por lo cual se le conoce como *dualismo de propiedades*) . Al referirme a la *naturaleza de las propiedades* intento señalar que este dualismo pretende llevar al terreno de la discusión filosófica el significado y uso de los conceptos utilizados en diversas disciplinas para describir y calificar a hechos relacionados con la mente. Por ejemplo, las propiedades mentales consideradas en la neurofisiología son distintas a las propiedades mentales según el conductismo, el psicoanálisis y la fenomenología. Una vez diferenciadas dichas propiedades, otra tarea del dualismo de propiedades consiste en situar explícitamente los diversos puntos de vista según la naturaleza de cada disciplina y buscar una forma adecuada de interacción entre ellos (Robinson,

2003). El resultado de la diversidad de enfoques que plantea el dualismo de propiedades ha dado como resultado que, en sus diversas variantes, sea una de las concepciones del problema de la conciencia más extendida y defendida en la actualidad.

Dentro el ámbito del dualismo de propiedades se pueden distinguir dos maneras de concebir las propiedades asociadas a la conciencia. Por un lado se encuentran las teorías *dualistas de propiedades emergentes*, las cuales postulan que dichas propiedades son el resultado de una compleja organización de sus componentes físicos, pero dicho resultado rebasa las causas físicas, por lo cual tales propiedades no son predecibles en términos de su naturaleza física[4]. Este tema es objeto de un debate muy extendido en diversas áreas de la filosofía, muy cercano al concepto conocido como *superveniencia* (Kim, 1998).

A diferencia de lo anterior, las teorías *dualistas de propiedades fundamentales* proponen que las propiedades de la conciencia no emergen de las propiedades físicas, sino que son componentes básicos de la realidad al igual que las propiedades físicas (Chalmers, 1996). Al ser ambas propiedades parte de la realidad, es posible una interacción causal entre ellas aunque su existencia sea independiente. Por otro lado el *panpsiquismo* afirma que todos los elementos que constituyen la realidad tienen propiedades psíquicas de distintos grados, y es en ese grado en que se encuentra el nivel de conciencia del objeto poseedor (Woodward, 1972).

En el otro extremo de la discusión encontramos a las teorías resultantes de la herencia de la tradición materialista, también conocidas como fisicalistas por basarse principalmente en conceptos de la física, las cuales conservan la idea inicial de la tradición de la que provienen, al explicar la conciencia como formas de funcionar de los procesos físicos. Si bien actualmente no existe una explicación completa, los apegados al

4 No todas las teorías emergentistas consideran que las propiedades que emergen rebasen la causalidad física. Sin ese rebase, las teorías emergentistas dejan de clasificarse dentro del dualismo forman parte de las teorías materialistas. Un ejemplos de lo anterior se puede encontrar en Bunge (2003).

materialismo afirman que el avance de la ciencia eventualmente llegará a ofrecerla (Churchland P. S., 1983).

Dentro del ámbito de las teorías fisicalistas encontramos al *eliminativismo*, el cual niega la existencia de la conciencia argumentando que el concepto mismo es el resultado de una manera equivocada de concebir la mente y que es deseable reformularlo o prescindir de su uso (Churchland P.S., 1983; Dennett, 1991). La teoría de la identidad postula que un estado mental es idéntico a un estado específico de ciertos componentes de sistema nervioso, lo cual se convierte en un problema al generalizar estados mentales mediante un concepto potencialmente proyectable a agentes no humanos (por ejemplo el dolor) (Lewis, 1972). Entonces la teoría de la identidad sin esa generalización problemática, postula que un estado único en cada individuo tiene un equivalente neuronal en el mismo (Davidson, 1980). A esta teoría se le conoce como la teoría de las *identidades particulares*.

Para el *funcionalismo*, un estado mental es una descripción única y óptima de un estado del mecanismo físico donde radica, pudiendo este ser o no biológico (Putnam, 1967). Dada la concepción anterior de funcionalismo, si se considera que las descripciones de los estados físicos son el resultado de una compleja organización de los componentes físicos básicos de un organismo o sistema, claramente nos encontramos ante la definición de propiedades emergentes. Sin embargo, a diferencia del dualismo, el fisicalismo no considera que dichas propiedades sean otra cosa que su descripción funcional. Desde el punto de vista anterior, es posible considerar que el funcionalismo es una especie de fisicalismo emergentista.

Por último el *fisicalismo no reductivo* apela a que los conceptos físicos utilizados en las teorías fisicalistas no son los adecuados para el análisis de las teorías de la conciencia, puesto que ambos se encuentran en distintos niveles conceptuales (Boyd, 1980).

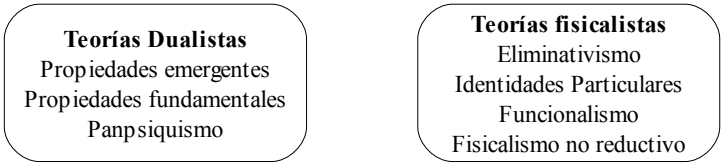


Ilustración 1: Clasificación de las teorías de la conciencia según su tradición filosófica.

Las teorías anteriores se presentaron conforme a la tradición de la filosofía de la mente a la que pertenecen, sin embargo existen postulados y presupuestos compartidos por las teorías en diversos niveles que las relaciona entre ellas de distinta forma según el criterio con que se evalúen. A continuación propongo otra posible clasificación.

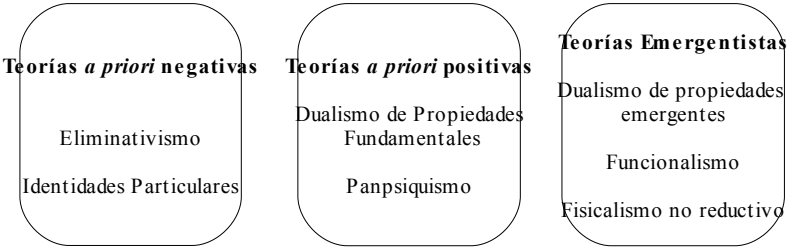
1.4.2 Clasificación según su estado ontológico.

Algunas de las teorías expuestas presuponen un estado ontológico de la conciencia previo a cualquier aseveración de la misma, proponiendo ya sea que no existe tal cosa como la conciencia, que la conciencia existe independientemente de otros elementos o que la conciencia emerge de la manera de organizarse de los elementos básicos de un sistema.

La primera opción es considerar la (no) existencia de la conciencia como ente en el mundo, es decir, su estado ontológico *a priori*. En su forma negativa, este presupuesto afirma que no existe tal cosa que reconocemos como conciencia, sino que es un concepto vago y equivocado. En su forma positiva, asevera que la conciencia es un ente existente en el mundo y que no puede ser reducido a conceptos más básicos. En este ámbito, el eliminativismo procede de manera negativa afirmado la inexistencia y argumentando la naturaleza equivocada del concepto y la teoría de la identidad si bien no reduce el concepto, lo hace idéntico a un estado físico, abriendo la posibilidad de prescindir de él. Por otra parte, la teoría de propiedades fundamentales y el panpsiquismo consideran *algo* ya existente en el mundo llamado conciencia que no requiere mecanismos previos a su existencia.

La contraparte del anterior es la noción de emergencia, pues en lugar de afirmar o negar una ontología *a priori*, considera que lo que se quiere capturar mediante el concepto de conciencia es una manera de concebir la interacción compleja de mecanismo más básicos y sus consecuencias. Es decir, no hay algo que exista en el mundo que podamos caracterizar como conciencia, sino que es el producto de la interacción de diversos sistemas complejos: es un fenómeno que emerge de esos sistemas.

Las teorías emergentistas ya mencionadas (el dualismo de propiedades emergentes y el funcionalismo), son las que asumen dicho presupuesto.



Ilusración 2: Clasificación de las teorías de la conciencia según su estado ontológico

Si bien reitero que el análisis anterior de las teorías de la conciencia no pretende abarcarlas todas, considero que las presentadas aquí son lo suficientemente diversas y generales como para dar una imagen del panorama actual y contextualizar el ámbito donde se desenvuelve la teoría de la conciencia concerniente a este texto.

2. La Inteligencia Artificial

En el capítulo anterior se mencionó que gran parte de la dificultad del abordaje al problema de la conciencia (y probablemente mucho de su encanto) se debe a su naturaleza multidisciplinaria. El viejo y obscuro cuestionamiento metafísico ahora puede dar luz y contribuir al desarrollo de áreas como la neurología y la psicología, pero ¿qué relación puede tener con el área de la computación?

Dado que el problema de la conciencia emerge, entre otros dilemas, como una pregunta acerca del pensamiento humano, es de sentido común considerar que sean las disciplinas relacionadas con el estudio del cerebro y los procesos mentales las encargadas de abordarlo. Sin embargo, paralelo al problema en cuestión, ha existido otro de al menos la misma longevidad e importancia: el hecho de considerar que los humanos (y en general los seres vivos) somos una especie de máquinas biológicas (La Mettrie, 1748). La idea afirmativa de lo anterior ha contribuido al descubrimiento de leyes y correlaciones en disciplinas como la medicina. Es esa misma idea la que llevaría a creer que en principio seríamos capaces de capturar el mecanismo del pensamiento humano mediante algún tipo de modelo formal a la usanza de las ciencias exactas.

El objetivo de este capítulo es recapitular someramente algunos de los momentos más importantes en el desarrollo de la computación tanto teórica como aplicada, enfatizando que en esa historia siempre existió una motivación ligada a la creación de un modelo que capture el pensamiento humano, para entonces definir lo que se pretende abarcar mediante el concepto de inteligencia artificial y así poner en la mesa las opciones filosóficamente verosímiles y actualmente en disputa en esta disciplina.

2.1 Los visionarios

Es probable que la idea de emular el pensamiento humano mediante un modelo matemático se remonte a un momento lejano en la historia de las matemáticas, sin embargo es en el siglo XVII donde es posible encontrar un

punto de referencia para ese proyecto ligado al área de la computación. El filósofo y matemático Gottfried Leibniz desarrolla una calculadora mecánica que era capaz de realizar las operaciones de suma, resta, multiplicación y división, conocida como *la rueda de Leibniz*. La relevancia del invento anterior es que le sirve de inspiración para especular y prever con atino algunos alcances y restricciones de la computación. Es de particular interés su trabajo enfocado al desarrollo de un lenguaje universal (*lingua characteristica*) que, entre otras cosas, sería la base a utilizar en su proyecto de desarrollar un cálculo formal que capturara la razón (Hernández, 1999). En palabras de Davis (1987, p. 136), *Leibniz esperaba mecanizar gran parte del pensamiento liberando a la mente de pensar en las cosas directamente en sí mismas*, en donde puede leerse su apuesta por la parte sintáctica sobre la semántica.

De lo anterior cabe resaltar que el cálculo de la razón y el lenguaje universal no fueron desarrollados nunca, por lo cual se puede considerar ese proyecto como un fracaso. Sin embargo, es admirable la claridad y familiaridad con la que Leibniz se refiere al pensamiento como un fenómeno formalizable en las matemáticas. Y es en esa familiaridad donde se puede observar la aguda intuición de Leibniz, puesto que el postulado de la mecanización del pensamiento y la mera construcción de su calculadora, no hubieran aportado elementos novedosos a la discusión filosófica[1]. Por el contrario, el hecho de describir el mecanismo como un sistema de manipulación sintáctica de símbolos a la manera de su famosa rueda, lo convierte en el punto de referencia histórico cuando se trata del desarrollo de la inteligencia artificial.

Dos siglos después que Leibniz y sin una mejora relevante en cuanto a recursos teóricos y técnicos, Charles Babbage continuó el desarrollo

1 Schickard y Pascal diseñaron y construyeron máquinas calculadoras con capacidad de sumar y restar antes que Leibniz. Sin embargo, es la rueda de Leibniz el punto de inicio de este análisis porque, más allá de un trabajo ingenieril, el dispositivo que diseñó Leibniz junto con el proyecto de la *lingua característica* documentan explícitamente la especulación de la formalización de todo pensamiento humano (no sólo lo relevante al cálculo aritmético), elemento que no se encuentra en los proyectos de Schickard y Pascal.

de las calculadoras mecánicas. A diferencia de Leibniz, la *máquina analítica* que diseñó nunca pudo ser construida, y las especulaciones de su alcance siempre se mantuvieron dentro del dominio de las matemáticas. Sin embargo, en el diseño del mecanismo de su máquina, Babbage planteó una serie de operaciones básicas y anticipó la existencia de un cierto tipo de problemas matemáticos que no se podrían resolver con ella. Según Gandy (1995, pp. 52-53), lo anterior tiene una fuerte semejanza con lo que después se conocería como funciones recursivas primitivas y el problema de la decisión. En otras palabras, a través de su máquina analítica, Babbage predijo con alguna precisión los límites que tendrían las computadoras digitales, a pesar de que el diseño de estas últimas no fue influenciado directamente por el de su máquina analítica.

2.2 Los modelos lógico-matemáticos

Para encontrar el origen de la computación como herramienta teórica de problemas de naturaleza matemática, es necesario tomar como un punto crucial el programa de Hilbert de principios del siglo XX conocido como *formalismo*, en el cual se pretendía fundamentar el conocimiento matemático mediante una serie finita de axiomas lógicos, correctos y consistentes. De esa manera, se pretendía que el pensamiento matemático fuera capturado y reducido mediante un modelo de manipulación sintáctica de símbolos. Así, como parte del programa anterior, Hilbert enuncia un problema (conocido como el décimo problema de Hilbert) donde se plantea encontrar un *procedimiento efectivo* (Hernández-Quiroz y Morado, 2006) para determinar si una ecuación diofantina tiene solución o no la tiene. Entonces, cuando los problemas del álgebra son transmitidos al cálculo lógico de primer orden, el problema anterior anuncia la naturaleza de un tipo de problemas más general, conocido como el *entscheidungsproblem* o el problema de la decisión: encontrar un algoritmo que permita decidir cuáles fórmulas (del cálculo de primer orden) son teoremas (Davis, 1987) .

En busca de una solución al problema de la decisión, algunos de los seguidores de Hilbert mantenían la esperanza de que pudiera ser resuelto,

mientras que otros creían que no se podría (Gandy, 1995). Es con la aparición del teorema de incompletitud del matemático Kurt Gödel (1929) cuando se establece una postura generalizada acerca del problema de la decisión: es improbable la existencia de tal algoritmo. Dicho teorema afirma lo siguiente: En cualquier formalización consistente de las matemáticas que sea lo bastante fuerte para definir el concepto de números naturales, se puede construir una afirmación que ni se puede demostrar ni se puede refutar dentro de ese sistema. La expresión anterior es conocida como el primer teorema de incompletitud, de la cual se deduce el segundo teorema de incompletitud que afirma lo siguiente: ningún sistema consistente se puede usar para demostrarse a sí mismo (Davis, 1987).

El análisis y tratamiento del concepto de *algoritmo* consecuencia del de *procedimiento efectivo* enunciado en el décimo problema de Hilbert y ligado a la solución del problema de la decisión, es lo que da pie a la creación de una serie de modelos distintos que intentan capturar la idea de lo que se requiere para responder dichos problemas (Hernández-Quiroz y Morado, 2006). Uno de los modelos de este tipo se generó en el trabajos del mismo Gödel y del francés Herbrand, donde se utiliza una caracterización muy precisa de las funciones recursivas para definir dicho procedimiento, mientras que por otro lado, Church desarrolla el modelo conocido como cálculo lambda, ambos modelos muy abstractos (Sieg, 2008). Otro modelo relevante es el de la máquina de Post, postulado por el matemático de ese apellido, en el cual el modelo comienza a acercarse a una realización mecánica en lugar de la completa abstracción (Uspenski, 1979). Sin embargo, es un modelo muy similar al anterior el que llama la atención del gremio debido a que todos parecen coincidir en la idea que el modelo capturaba la idea de procedimiento efectivo: la máquina de Turing, propuesta por Alan Turing. Si bien el mismo Turing posteriormente demostró la equivalencia su modelo y el cálculo lambda propuesto por Church, las características de la máquina de Turing llevaron a un acuerdo unánime (Sieg, 2006). Gandy (1995, pp. 93) comenta al respecto:

> Lo que Turing hizo fue mostrar que el cálculo podía dividirse en la iteración (controlada por un programa) de operaciones concretas extremadamente simples; tan concretas que podían ser descritas fácilmente en términos de mecanismos (físicos). (Las operaciones del cálculo lambda son mucho mas abstractas).

Cabe mencionar que, si bien es probable que inicialmente la máquina de Turing haya sido concebida como una herramienta lógica para abordar el problema de la decisión, conforme se descubría su capacidad de cómputo, las expectativas de Turing se incrementaron (o al menos fueron expresadas explícitamente) hasta el punto de tomar una posición con respecto dicho mecanismo y su relación con el pensamiento humano. En el texto *Computing Machinery and Intelligence*, Turing (1950) aborda la pregunta acerca de la capacidad de las máquinas para pensar. Allí propone una forma de evaluar dicha pregunta (conocida como la prueba de Turing[2]) y expone su posición al respecto. Su respuesta es cauta y explícitamente limitada, pero parece evidente que su apuesta es a favor de una respuesta afirmativa, confiando en que los avances tecnológicos le darían la razón.

A partir de la información anterior, pretendo resaltar que la idea de generar un modelo formal capaz de capturar (o al menos simular) el pensamiento humano subyace en la mayoría los proyectos lógico-matemáticos anteriormente mencionados, ya sea de manera explícita o implícita. Es entonces posible pensar que, si la computadora digital es la implementación física de los modelos lógico-matemáticos abstractos, entonces funciona como plataforma para llevar a la práctica las teorías relacionadas con la computación y el pensamiento humano.

La afirmación anterior es polémica en varios sentidos. El primero de ellos es referente a la relación entre los modelos lógico-matemáticos y el

2 La prueba de Turing consiste en un juego de imitación, donde un interrogador humano realiza preguntas a dos personas que no puede ver, con la finalidad de adivinar su género. Cuando alguno de los sujetos es reemplazado por una computadora y el interrogador no puede distinguir con certeza cual humano fue sustituido, entonces se considera que la prueba fue exitosa.

desarrollo de la computadora. Si bien parece innegable una relación estrecha entre ambos, no es posible afirmar con certeza que los primeros sean necesariamente indispensables para la existencia de la segunda (Davis, 1987). Otro punto a discusión es si la idea de modelar el pensamiento humano efectivamente era tan constante como se afirma. Para abordar lo anterior, es necesario exponer la posición al respecto de uno de los más involucrados en el desarrollo de la computadora digital: John Von Neumann. Para resaltar la relevancia de este personaje, basta recordar que uno de los reconocimientos mayores que se le acreditan (entre muchos otros), es el desarrollo del modelo de la computadora digital, a tal grado que dicho modelo (el cual sigue siendo la base de las computadoras digitales construidas al día de hoy) es conocido como la arquitectura[3] Von Neumann. Con respecto al primer punto polémico, autores como Davis (1987) consideran que la aportación de Von Neumann está fundada en gran parte en el trabajo de Turing y que los modelos lógicos necesariamente anteceden al desarrollo de la computadora digital. Sin embargo el debate continúa sin una definición categórica. Con respecto al segundo punto, Boden (1988, p. 2) ofrece un punto de referencia:

> A diferencia de McCulloch y Pitts, sin embargo, Von Neumann no creía que la lógica binaria pudiera modelar el pensamiento humano: en sus palabras "el lenguaje del cerebro no es el lenguaje de las matemáticas"; su sugerencia que la probabilidad termodinámica se adapta mejor ha sido revivida (Von Neumann, 1958).

en el cual es posible observar que la visión de Von Neumann explícitamente niega lo expuesto anteriormente.

3 La arquitectura de computadoras es el diseño conceptual y la estructura operacional fundamental de un sistema de computadora. Es decir, es un modelo y una descripción funcional de los requerimientos y las implementaciones de diseño para varias partes de una computadora, con especial interés en la forma en que la unidad central de proceso trabaja internamente y accede a las direcciones de memoria.

Lo relevante al respecto es que aún aceptando que Von Neumann y sus seguidores creyeran que la computadora digital que desarrollaban era ajena a algún proyecto del modelado de la mente humana, su existencia abrió el panorama para este tipo de visión. Así vemos la mención en la cita anterior del psicólogo y psiquiatra W. S. McCulloch y el matemático W. H. Pitts, quienes comparan la estructura lógica en los circuitos de las computadoras digitales con conjuntos de neuronas interconectadas (McCulloch y Pitts, 1943). Igualando los estados binarios de encendido y apagado de la computadora con las propiedades de la conducción nerviosa, afirman que las propiedades lógicas del cerebro como un todo pueden ser comprendidas en términos de las propiedades lógicas de las células que lo constituyen. Enfocándose en el potencial computacional de las neuronas más que en su fisiología real, argumentan que el tipo específico de unidad neuronal, funcionando bajo restricciones particulares, tendría propiedades lógicas específicas que podrían ser modelables.

Es a partir de entonces, con las ideas concretas de McCulloch y Pitts y la disponibilidad de la computadora digital, que la vieja intuición del modelado de pensamiento toma un rumbo más orientado a la ciencia empírica, dando paso al concepto de *inteligencia artificial*, ya con la intención de contar con un sustento que pueda definir el rumbo de los proyectos relacionados. Sin embargo, esto lejos de unificar las opiniones al respecto, ha generado nuevas y mayores polémicas, algunas de las cuales presento a continuación.

2.3 La inteligencia artificial fuerte

Las ideas expuestas en las secciones anteriores de este capítulo intentan describir el contexto histórico de la ciencia computacional, relatando brevemente el desarrollo de algunos problemas, ideas y herramientas que al converger en un momento específico, parecerían proponer la línea a seguir de una nueva disciplina. El objetivo de lo anterior es intentar que el lector se forme un criterio y una idea de los objetivos y métodos asociados a la

inteligencia artificial, para así proceder con el problema de su definición y delimitación.

Comúnmente, los no especialistas en el área, asocian la idea de inteligencia artificial (abreviada IA) con niños-robot a la usanza del cine, lo cual no está del todo equivocado, pues existe una relación cercana entre la robótica y la IA como se menciona al posteriormente en este capítulo. De forma paralela, las personas relacionadas con la industria piensan en maquinaria automatizada que reemplaza la mano de obra. Algunos neurocientíficos creen que es una analogía para la comprensión del funcionamiento de algunos mecanismos biológicos, pero nada más que eso. Dentro del gremio de los informáticos, algunos creen que la IA se reduce a programas que utilizan inferencia estadística y probabilística para generar *nuevo conocimiento*, como los llamados *sistemas expertos* y *programas de minería de datos*.

Este tipo de acepciones son consideradas por Winston (1992, p. 6), como IA de *objetivo ingenieril*, el cual abarca tanto el área de los programas de software, como los mecanismos de control ligados al movimiento autónomo (robótica). Por otro lado, aquello que se asocie con el problema del modelado del pensamiento humano a un nivel conceptual y abstracto, lo llama IA de *objetivo científico*.

Es importante resaltar que las diversas concepciones de la IA han dado pie a una ambigüedad con respecto al uso del concepto, por lo que algunos autores prefieren acuñar nuevos términos para referirse al área de la IA relevante al pensamiento. Por ejemplo, Boden (1988) se refiere a tal concepto como *psicología computacional*, mientras que Jordan y Russell (1999) prefieren usar el término *inteligencia computacional*.

Dado lo anterior, propongo que la parte relevante al análisis filosófico de la IA será la relacionada al objetivo científico y es a lo que, a partir de aquí, me referiré como IA. Cabe mencionar que la distinción anterior no establece la frontera precisa que originalmente sería deseable,

pues existen autores que consideran que la descripción de los mecanismos de control motor están intrínsecamente ligados con el pensamiento de nivel abstracto. Si se acepta el postulado anterior, áreas como la robótica (Brooks, 1986) y el estudio de la visión por computadora (Marr, 1982) podrían considerarse partes intrínsecas de la IA como deseamos identificarla. Esto no afecta al objetivo de este texto siempre y cuando quede claro que lo relevante aquí es que el modelado del pensamiento al cual nos referimos debe permitir niveles de complejidad suficientes como para que los modelos generados puedan dar cabida a la idea de la conciencia presentada en el capítulo anterior, y si estos modelos se ocupan o no de aspectos motores, no es objeto de este análisis.

Ya dentro de las fronteras establecidas, es posible distinguir los dos puntos de vista generales relacionados con la IA relevantes a este texto, los cuales son breve y claramente descritos en el polémico texto de Searle (1980, p. 67) llamado *Minds, Machines and Programs*:

> Encuentro útil distinguir lo que llamaré la IA (Inteligencia Artificial) "fuerte" de la IA "débil". De acuerdo a la IA débil, el principal valor de la computadora en el estudio de la mente es que nos provee de una herramienta muy poderosa. Por ejemplo, nos permite formular hipótesis de pruebas en una manera más rigurosa y precisa. Pero de acuerdo con la IA fuerte, la computadora no es solamente una herramienta en el estudio de la mente; por el contrario, la computadora apropiadamente programada es una mente, en el sentido que las computadoras, dados los programas correctos, puede decirse que literalmente entienden y tienen otros estados cognitivos.

La relevancia de resaltar la distinción anterior es que, a la luz del contexto histórico aquí presentado, pareciera posible trazar una línea de ideas que inician con Leibniz y su proyecto con respecto a la mecanización del pensamiento; continuando la idea de formalización de Hilbert en el área

matemática y posteriormente con la proyección de la máquina de Turing hacia propósitos muy generales como la simulación de la conducta verbal humana, para terminar con la concepción computacional del sistema nervioso de McCulloch y Pitts. De esta forma, al extrapolar esta línea hasta nuestros días, ya en presencia de la computadora digital como la herramienta prevista y soñada por los iniciadores de esta tradición, parecería que las ideas de nuestros visionarios en cuestión estarían, si no comprometidas, al menos cercanas y empáticas con respecto al proyecto de la IA fuerte.

2.4 Consideraciones filosóficas acerca de la IA fuerte

Una primera aproximación filosófica hacia el proyecto de la IA fuerte, comienza con la polémica de su posible existencia. De hecho, el autor del término (Searle, 1980) lo utiliza para argumentar en contra, señalando la imposibilidad de la semántica en las computadoras, y su inherente existencia en el pensamiento humano. Si bien es cierto que el argumento en sí tiene muchos presupuestos y postulados polémicos (y probablemente erróneos) que han dado pie a un gran número de esquemas para confrontarlo (Preston y Bishop, 2002) y a un sinnúmero de réplicas (Cole, 2008), también constituye un punto de convergencia al señalar aspectos de la concepción de la IA fuerte no previstos anteriormente que deben ser precisados. Es en el proceso de esta refinación de conceptos donde es posible encontrar el origen de la discusión filosófica actual presentada a continuación.

2.4.1 Representacionalismo.

Gracias a los señalamientos de Searle, se puede distinguir una primera y radical lectura del proyecto de la IA fuerte considerada como una teoría de la identidad, donde se asume que el cerebro *es* (funciona y debe ser explicado como) una computadora de arquitectura Von Neumann: un dispositivo de estados discretos que almacena representaciones simbólicas y las manipula de acuerdo a reglas sintácticas; además, los pensamientos son

representaciones simbólicas y los procesos mentales son secuencias causales generadas mediante un conjunto de reglas. Esta tesis que identifica al cerebro con una computadora, debilitada mediante algunas variantes y sutilezas, es reconocida como *representacionalismo*[4].

Dado que computadora es un mecanismo diseñado para manipular símbolos basado en la lógica, al identificar al cerebro con una computadora, el representacionalismo acarrea las ventajas y desventajas de un sistema de lógica simbólica. En primera instancia, la asignación de significados a símbolos (posiblemente arbitrarios) y un conjunto de reglas correctas, permite preservar las relaciones semánticas deseadas (como lo establece el teorema de completitud de Gödel[5]). En términos de Clark (1996, p. 5):

> Somos extraordinariamente bien portados semánticamente. Con frecuencia disfrutamos *flujos de pensamiento (trains of thought)* que tengan sentido racionalmente. La explicación más simple de todo esto es que razonamos utilizando un sistema de símbolos y un conjunto de reglas de transformación semánticamente sensibles.

Lo anterior implica, de una manera análoga a la maquina de Turing y a la computadora digital, que los problemas asociados a la computabilidad y al problema de la decisión están inmersos también en la mente humana.

---

4 En algunas teorías, el término *representacionalismo* tiene un uso distinto al que se presenta aquí. En todo caso, todas las concepciones aceptarían la existencia de un *objeto intencional*. El compromiso de ese objeto intencional con el enfoque simbólico - computacional es lo que hará inaceptable esta definición en algunos casos. La forma en que se presentó el concepto de *representacionalismo* aquí es referida de distintas maneras por algunos autores, por ejemplo: Varela, Thompson y Rosch (1992. p. 64) lo llaman *cognitivismo*, Clark (1996, p. 3) lo llama *IA de sistemas simbólicos (symbol-system AI)* y Boden (1988, p. 229) lo refiere como *formalismo*.

5 El teorema de completitud de Gödel establece en su forma más conocida que, en una lógica de primer orden, toda fórmula que es válida en un sentido lógico, es demostrable. Esto es, que, para cada fórmula válida, existe una lista finita de pasos en los que cada paso o bien invoca a un axioma o es obtenido a partir de pasos previos mediante una básica de regla de inferencia.

Otro aspecto a tomar en cuenta para el representacionalismo, es que, de manera relativamente fácil, es posible explicar la forma sistemática y la potencialmente infinita productividad del pensamiento a partir de un conjunto finito de recursos iniciales (símbolos o representaciones atómicas) y un conjunto de operaciones mentales iterables, pues esta combinación produciría, en principio, una infinidad de pensamientos a partir de una base finita de materiales (Fodor y Pylishyn,1988) .

Hasta ahora, lo expuesto acerca del representacionalismo parece presentar más ventajas que desventajas como postura de la IA, pues el tránsito del cerebro a la computadora digital sería prácticamente transparente. Sin embargo, gran parte del peso de la teoría radica en especulaciones por comprobar o refutar a la larga por la neurociencia. Otro aspecto problemático radica en que el representacionalismo parece no tener una explicación obvia y *natural* para el tipo de conocimiento asociado a habilidades, tales como el control de la locomoción o el reconocimiento de imágenes (Churchland, 1989), pues es difícil capturar esos mecanismos mediante el uso de un proceso de símbolos en serie como lo haría una computadora Von Neumann.

2.4.2 Conexionismo

En la sección 2.2 se menciona a McCulloch y Pitts como pioneros en los modelos computacionales motivados por la estructura de un sistema nervioso biológico. Dichos modelos planteaban que la manera de interconectar los circuitos sería la vía para aproximarse al modelo neuronal, pero estaba muy lejos de tener una aplicación práctica. Posteriormente, Rumelhart y McClelland (1987a) redefinen el modelo teórico anterior al cambiar la idea de comparar el hardware con el sistema nervioso por un modelo de simulación mediante técnicas de programación en el desarrollo software, que llegará a conocerse como *redes neuronales artificiales* (Zilouchian, 2001).

Un enfoque de las redes neuronales artificiales enfatiza que algunos de los procesos biológicos y psicológicos se explican de forma mucho más

natural si se considera que el procesamiento neuronal se efectúa de forma paralela (Rumelhart y McClelland, 1987b), al contrario de la forma serial planteada por el representacionalismo. Esta idea conocida como *procesamiento distribuido en paralelo* recibió un fuerte soporte empírico gracias a los resultados exitosos de las aplicaciones de las redes neuronales artificiales en diversas disciplinas (Jamshidi y Zilouchian, 2001), dando así un fuerte empuje a una nueva opción filosófica en la IA conocida como *conexionismo*.

Las principales diferencias entre los modelos representacionalista y conexionista abarcan tres aspectos básicos: la forma de representar el conocimiento, las operaciones básicas de procesamiento y el uso intrínseco de un conjunto de algoritmos poderosos. El conexionismo niega que el conocimiento deba representarse mediante símbolos atómicos concatenados, proponiendo que tal representación se realice a través de patrones de activación entre grandes cantidades de neuronas idealizadas (pequeñas unidades de procesamiento) que codifican contenidos específicos (Clark, 1996, p. 7). Los patrones de activación mencionados se representan en este modelo mediante vectores numéricos, por lo cual las operaciones básicas son cálculos matemáticos de estos vectores. La semántica está inmersa en la estructura de cada patrón, por lo cual no hay símbolos arbitrarios que puedan carecer de ella. Un ejemplo de lo anterior es la forma de determinar el grado de similitud entre dos patrones, realizada por una función matemática aplicada a los vectores asociados a los patrones, lo cual puede leerse como una comparación casi directa entre los contenidos semánticos de cada uno de ellos. El caso de la postulación de un conjunto de algoritmos poderosos responde a la necesidad de dar cuenta del proceso de aprendizaje, pues a diferencia del representacionalismo, aquí no se cuenta con una base fija ni con operaciones iterantes. La manera de proceder entonces es postular un tipo de *entrenamiento*, donde inicialmente el contenido de los vectores numéricos es aleatorio y se va modificando de una forma heurística, pues el agente en cuestión es expuesto a una serie de datos de entradas y sus respectivas salidas deseadas, mediante lo cual se van alterando los contenidos de los vectores y

así se van perfeccionando los patrones de activación. Un ejemplo que casa muy bien con el postulado anterior es el aprendizaje del lenguaje (Rumelhart y McClelland, 1987c).

Así, al no depender de concatenación de símbolos, el conexionismo permite representaciones distribuidas y de sistemas complejos sin la necesidad de aislar a cada elemento con una representación propia, mejorando también la representación de contextos y relaciones complejas. A diferencia del representacionalismo, en esta teoría los contenidos semánticos están intrínsecamente ligados a cada patrón, reivindicando así la importancia de la semántica. Así mismo, puede considerarse que la capacidad de *entrenamiento* es tan poderosa como para modelar casi cualquier función coherentemente imaginada (Clark, 1996, p. 10), igualando en este aspecto el poder de las reglas sintácticas. A pesar de todas estas ventajas, la mayoría de las críticas a esta posición apuntan a que el nivel de abstracción de este modelo es muy alto, que este tipo de procesamiento es muy costoso para un sistema biológico y es difícil imaginar que de hecho funcione así. Otro aspecto problemático es la presuposición de una única y excesiva localización funcional[6], ya que los patrones de activación están asociados a una estructura fija e idealizada de neuronas, lo cual causa conflicto con los datos de la neurociencia asociados al carácter distribuido de mecanismo mentales y a la plasticidad de la mente.

6 Con *localización funcional* me refiero al mecanismo subyacente necesario para efectuar las funciones matemáticas o algorítmicas del modelo formal, no a las funciones biológicas de los organismos.

2.4.3 Tendencias naturalistas: vida artificial y sistemas basados en conducta

Las dos teorías anteriores descritas de manera muy general, son las posiciones dominantes al respecto de la filosofía de la IA y donde se ha centrado el debate de los últimos años; sin embargo no son las únicas alternativas. Con las descripciones anteriores es posible darse cuenta de que ambas teorías tienen un punto débil en común: son producto de abstracciones provenientes de problemas de naturaleza lógica-matemática y con ello intentan responder a los cuestionamientos que emergen de las áreas de la neurología, la biología y ciencias afines, por lo cual, ante los ojos de los científicos cercanos a estas áreas, estos postulados parecen rebuscados y ajenos al funcionamiento *real* del que puede dar cuenta la ciencia empírica. De allí es que nacen propuestas como la vida artificial simulada y el de los sistemas basados en conducta.

La vida artificial busca capturar la lógica de los sistemas vivos mediante la simulación mediante el uso de software. El objetivo es estudiar el fenómeno de los sistemas vivos con la finalidad de entender el complejo proceso de información que los define, usando la simulación computacional. Se simulan los elementos relevantes de un ecosistema y los agentes que interactúan con él. De esa forma, la detección y manipulación de variables relevantes permite explicar muchos aspectos biológicos y evolutivos. Otra meta del proyecto de vida artificial es capturar algunas propiedades emergentes de las *sociedades* formadas por la interacción de los agentes (Langston, 1995). Una herramienta muy utilizada en este tipo de programas, son los llamados *autómatas celulares*, los cuales son modelos lógicos de sistemas dinámicos que evolucionan en pasos discretos y constan de una colección masiva de objetos simples que interactúan localmente unos con otros (Wolfram, 1994).

Por otra parte, un enfoque nacido del diseño y construcción de robots autónomos es el que se basa en la idea realizar componentes sencillos para el control de un mecanismos específicos (como control motor de un

único miembro). Una manera de interactuar de esos componentes será lo que dé origen a los *sistemas basados en conducta* (Brooks, 1986). En ellos se considera que cualquier forma de abordar el desarrollo de la inteligencia que radique en alguna forma de representación (tanto representacionalismo como conexionismo), aspira a metas muy limitadas. La propuesta es abordar la inteligencia de como un mecanismo de control que incrementa su complejidad dependiendo estrictamente de la percepción y la acción que pueda tener el ente (organismo o robot) con el mundo (Brooks, 1991). Para efectuar esta manera incremental de abordar la inteligencia, es necesario construir agentes que operen en ambientes dinámicos, de tal manera que no contengan mapas extensos del mundo, sino que actúen coherentemente por la integración de sus componentes y que sean capaces de incorporar información de su medio y actuar en consecuencia.

Así como el modelo teórico del conexionismo nace de una idealización de las redes neuronales artificiales, los sistemas basados en conducta proviene de la idea de no restringir la conciencia a un sitio particular del cuerpo, sino a una continuidad entre el procesamiento cognitivo interno y la información externa del mundo físico. Este tipo de teorías que incorporan a la conducta y al entorno se les conoce como *cognición situada,* y sus postulados provienen en su mayoría de la biología y se fundamentan fuertemente en la fenomenología. La teoría de la enacción es uno de los representantes más conocidos (Varela, Thompson y Rosch, 1992). Díaz (2007, p. 122) considera que una de las aplicaciones más claras de esta posición concierne al problema mente-cuerpo en el sentido de que se integra al comportamiento en la esfera de las actividades mentales o cerebrales, estableciendo con ellas un proceso unitario de múltiples manifestaciones.

Hasta aquí la breve recapitulación del universo donde se desenvuelven las teorías filosóficas de la IA con sus respectivos fundamentos en las ciencias cognitivas, donde es posible reconocer tres grandes etapas por las que han pasado tanto la IA basándose siempre en los fundamentos que le proveen las ciencias cognitivas. En primer lugar y como punto de partida se

encuentra el representacionalismo que asume un mecanismo cerebral como una máquina de arquitectura Von Neumann (o una máquina de Turing, pues sus limites teóricos son equivalentes). La segunda etapa la representa el conexionismo con base en las posibilidades que proveen las redes neuronales y por último, la etapa más reciente la marcan la vida artificial simulada y los sistemas basados en conducta, donde el entorno y la conducta cobran importancia gracias a los postulados de la cognición situada.

Todo lo expuesto aquí tiene como finalidad proporcionar al lector las herramientas y referencias que le permitan entender, ubicar y eventualmente evaluar la propuesta de Chalmers con respecto a la forma en que su teoría de la conciencia es congruente con el proyecto de la IA fuerte.

3. La Teoría de la Conciencia de Chalmers

Con lo expuesto en los capítulos anteriores, pretendo ofrecer al lector un punto de entrada a los temas que conciernen al análisis de la teoría de la conciencia de Chalmers presentado aquí, así como una referencia bibliográfica para profundizar al respecto. De esta manera, el objetivo del presente capítulo será ofrecer un resumen del argumento y de los postulados principales de la propuesta de Chalmers para abordar el problema de la conciencia , con la finalidad de invitar al lector a realizar una lectura crítica utilizando las herramientas obtenidas hasta el momento y sus ideas propias. Así, al finalizar esta sección, propondré una ubicación de esta teoría en el marco desarrollado en el capítulo primero, resaltando sus presupuestos ontológicos y sus compromisos teleológicos con respecto a la conciencia.

3.1 La conciencia y el reduccionismo

Chalmers considera que los esfuerzos para explicar la conciencia[1] apelando a la neurofisiología y otras teorías del tipo funcional son en principio infructuosos, pues por más que se conozca acerca del funcionamiento y de la estructura, la pregunta con respecto a la razón de la experiencia subjetiva sigue sin responderse. El autor lo expone de la siguiente forma (Chalmers, 1995, p. 18):

> Una vez que se encuentran los detalles del aspecto físico, las propiedades funcionales y estructurales aparecen como consecuencia automática, pero la estructura y dinámica de los procesos físicos lleva únicamente a más estructura y dinámica; entonces son estructura y funciones todo lo que podemos esperar que expliquen.

1 Como se estableció en el capítulo 1, con el término conciencia me refiero al aspecto fenoménico de la misma.

Para sustentar el argumento anterior que afirma que las propiedades de la conciencia superan a las propiedades físicas (o que no supervienen lógicamente a ellas)[2], Chalmers postula una serie de experimentos mentales, de los cuales menciono brevemente dos. El primero sugiere la posibilidad de la existencia de individuos físicamente idénticos a nosotros pero ausentes de conciencia (zombis fenoménicos), mientras que segundo considera la posibilidad de la existencia de *qualia invertidos*. Con ello, el autor intenta probar que las propiedades físicas en cada caso no son suficientes para dar cuenta de la experiencia consciente en los individuos.

Dado lo anterior, Chalmers propone intentar explicaciones no reductivas. Ahora, considerando que gran parte de la actividad científica es explicar los fenómenos del mundo de manera (parcialmente) reduccionista[3], si se desea que una teoría de la conciencia vaya más allá de la especulación metafísica, entonces la propuesta anterior parece implicar un gran reto en muchos aspectos, pues aunado a las expectativas planteadas en el capítulo 1, se debe esperar también que dicha teoría ofrezca una explicación no reduccionista y que sea (por lo menos) plausible científicamente en este momento, esperando que eventualmente sea compatible con la ciencia, y que a la larga también llegue a formar parte ella.

3.2 Dualismo naturalista

Las condiciones enumeradas hasta el momento definen tanto el rumbo como la estructura genérica o *molde* de lo que Chalmers considera que debería cumplir cualquier teoría de la conciencia. Ya entrando en el terreno constructivo, la propuesta del autor denominada *dualismo naturalista*, es una variedad del dualismo de propiedades, donde la propuesta es expandir la

2 En el artículo *Facing up to the problem of consciousness* (Chalmers, 1995) la justificación de la idea de que la conciencia rebasa las propiedades físicas se queda en términos conceptuales y es hasta la publicación de su libro (Chalmers, 1996) donde dicha justificación apela a la *superveniencia lógica* en un análisis más profundo.

3 No pretendo entrar en el debate del reduccionismo en la ciencia, solamente intento recalcar que las explicaciones reduccionistas forman una parte importante en la actividad científica, al grado de que parece difícil imaginar dicha actividad sin tales explicaciones.

ontología del mundo al considerar la conciencia como una propiedad fundamental, es decir, no explicable en términos más simples; tal como sucede con el electromagnetismo en la teoría de Maxwell. Chalmers lo define de la siguiente forma (Chalmers, 1996, p. 125):

> El dualismo implicado aquí es una especie de dualismo de *propiedades*: la experiencia consciente involucra propiedades de un individuo que no están implicadas por las propiedades físicas de ese individuo, aunque puede depender nomológicamente de esas propiedades. La conciencia es una *característica* del mundo más allá de sus características físicas.

Así como está enunciada, la propuesta anterior parece no ser nada más que una especulación metafísica de las del tipo que se desea evitar debido a que, a primera vista, no parece tener un carácter científico. Sin embargo, ya podemos ver en la cita anterior una mención a una posible dependencia nomológica de las propiedades físicas por parte de las propiedades conscientes, donde se puede prever un intento por aterrizar la especulación. Es entonces que Chalmers genera una propuesta donde el funcionalismo forma una parte indispensable, siendo una fuente de información *objetiva* a la cual se propone *atar* el difícil tratamiento del problema intrínseco a la naturaleza subjetiva de la conciencia.

3.3 Principios psicofísicos

Desde mi punto de vista, es gracias a lo anterior que la teoría de la conciencia de Chalmers cobra mayor relevancia a nivel filosófico, pues con el objetivo de hacerla asequible desde un enfoque científico, propone que la ciencia utilice una serie de criterios distintos a la intersubjetividad que la caracteriza, pues la naturaleza solipsista de la conciencia parece no permitirlo. Algunos de esos criterios mencionados por el autor son: simplicidad, coherencia interna, coherencia con teorías de otras áreas, la

habilidad de reproducir las propiedades de la experiencia que nos son familiares para nuestro propio caso, y hasta el hecho de que coincidan con el sentido común. Para cumplir el requisito anterior, Chalmers requiere de postular lo que llama *principios psicofísicos,* los cuales describe de la siguiente manera (Chalmers, 1995, p. 22):

> Una teoría no reductiva de la conciencia consistirá en un número de *principios psicofísicos*, principios que conectan las propiedades de los procesos físicos con las propiedades de la experiencia[4], Podemos pensar estos principios como una forma de encapsular la manera en que la experiencia emerge de lo físico. Al final, estos principios deben decirnos qué tipo de sistemas físicos tendrán experiencias asociadas, y para los sistemas que las tengan, deben decirnos qué tipo de propiedades físicas son relevantes para la emergencia de la experiencia, y qué tipo de experiencia debemos esperar que produzca algún tipo de sistema.

Ya para cerrar su propuesta, Chalmers propone una serie de novedosos y controversiales principios psicofísicos, que si bien también son altamente especulativos, podrían eventualmente ser estudiados y refinados para evaluar su valor como fuente de información científicamente aceptable. A continuación resumo dichos principios.

3.3.1 El principio de la coherencia estructural

Si bien el dualismo naturalista recalca que las propiedades de la conciencia no están implicada por las propiedades físicas, también propone una dependencia nomológica entre ellas. Es esa dependencia lo que intenta capturar el principio de la coherencia estructural, afirmando que la

4 En el artículo (Chalmers, 1995), el autor utiliza el término *experiencia* como sinónimo de *conciencia*, dado que se refiere a la parte de la *experiencia subjetiva de la conciencia.*

experiencia consciente siempre va acompañada del proceso funcional definido por el autor como *percatación*[5] *(awareness)*: el acceso (a través del sistema nervioso central) a los contenidos informacionales que rigen (parcialmente) al comportamiento.

Debido a que el percatarse es un proceso funcional, es posible que cada sujeto pueda caracterizar y describir su propia percatación en términos cognitivos completamente funcionales sin tener que apelar a la experiencia subjetiva, por lo cual dichos términos podrán ser objetivos (y en principio científicos). Así, gracias a que existe un isomorfismo entre la conciencia y la percatación, es posible que otros sujetos sean capaces de reconstruir la experiencia consciente de otro utilizando como base la descripción de la percatación de éste último y las experiencias conscientes que han tenido en situaciones similares a la descrita.

La idea anterior pretende reflejar el hecho que la conciencia y la cognición no flotan libres una de la otra, sino que interactúan coherentemente de manera íntima. Sin embargo, Chalmers considera que la relación anterior termina en la incapacidad de la descripción funcional de dar cuenta del estado consciente sin apelar a la experiencia propia, argumentando de manera similar a Jackson (1982) que el conocimiento funcional total no provee la información suficiente para reconstruir una experiencia fenoménica, pues nunca se podrá explicar los *qualia* sin apelar a la experiencia subjetiva propia.

3.3.2 El principio de la invariancia organizacional

Chalmers propone como experimento mental que imaginemos que la neurociencia llega a ofrecer la posibilidad de empatar funcionalmente ciertos componentes básicos del sistema nervioso, y así poder ofrecer sustitutos artificiales; a la manera que funciona un marcapasos pero a nivel neuronal. Así, el reemplazo de un mecanismo neuronal de bajo nivel (como

5 Retomo el poco usual término de *percatación* de la versión traducida al castellano del libro de Chalmers (1996).

una única neurona) por uno causalmente isomorfo, no cambiaría la experiencia consciente del sujeto en cuestión, lo cual tampoco sucedería si el reemplazo se conservara en el mismo nivel de complejidad, aunque fuera de un número significativo de esos componentes o eventualmente todos. De lo anterior, el autor concluye que dos sistemas con la misma *organización funcional* en detalle (*fine-grained*)[6] tendrán experiencias cualitativamente idénticas. Entonces, lo que importaría para la emergencia de la conciencia no sería el sustrato físico, sino los patrones abstractos de la interacción causal de sus componentes.

Es importante recalcar que es muy relevante la concepción del término *organización funcional*, pues es lo que posteriormente se empleará para modelar lógicamente su propuesta en el área de la IA, y por ello presento la definición textual (Chalmers, 1996, 247):

> La organización funcional puede comprenderse mejor como el *patrón abstracto de interacción causal* entre las diversas partes de un sistema y, quizás, entre esas partes y las entradas y salidas externas. Una organización funcional se determina especificando 1) un número de componentes abstractos, 2) para cada componente, un número de estados posibles diferentes, y 3) un sistema de relaciones de dependencia que especifican cómo el estado de cada componente depende de los estados previos de todos los componentes y de las entradas en el sistema, y cómo las salidas del sistema dependen de los estados previos de los componentes.

3.3.3 El principio del doble aspecto de la información

Los principios psicofísicos mencionados hasta el momento, si bien pueden ofrecer una explicación de los mecanismos subyacentes en el problema de la conciencia, su ámbito explicativo se limita a una propuesta de

6 En la sección 4.1 argumento que es crucial especificar el nivel de detalle mencionado. Por el momento lo presento tal como lo hace el autor.

teoría de los sistemas y a ciertas afirmaciones con respecto a la neurociencia. Sin embargo, el título del libro donde se presenta esta teoría contiene la frase *en busca de una teoría fundamental*, objetivo que se vislumbra aún lejano apelando a los principios anteriores. Entonces se presenta la necesidad de enlazar la tesis ontológica de las partículas fundamentales de la conciencia (y el panpsiquismo resultante[7]) con los mecanismos fundamentales que puedan dar cuenta de los mecanismos de alto nivel expresados en los principios psicofísicos anteriores.

Para ello, Chalmers retoma el modelo de la teoría de la información postulado por Shannon (1948). A través de dicha propuesta, Shannon se propone describir el mecanismo del flujo de la información sin apelar a la semántica subyacente en ella. Lo anterior lo realiza mediante la descripción de un mecanismo general para toda la información que no se encuentre ligado a contextos específicos. Con ese objetivo en mente, Shannon postula una serie de conceptos completamente abstractos (sin referencias a objetos *reales*) tales como *espacio informacional* y *estados de información*. Con espacio informacional se pretende capturar la idea de que la información consiste en un *molde contenedor* de estructuras informacionales básicas (un espacio abstracto), y que cada manera de *llenar* ese molde es un estado informacional. Así, cada manera distinta de llenar el espacio informacional se reconocerá por la forma en que difiere de las demás, por lo cual la información será el conjunto de diferencias encontradas en determinado espacio informacional. El ejemplo más básico de espacio informacional es el de un bit, donde el molde consta de un solo espacio que puede contener uno de dos símbolos (cero o uno). En este ejemplo existen dos posibles estados de

7 En el principio del doble aspecto de la información, Chalmers defiende la idea de que las partículas fundamentales de la conciencia están ligadas a la información. Por tanto, al igual que la información, la conciencia es ubicua. Retomando el argumento de Nagel del *como es ser un murciélago*, afirma que un termostato tiene experiencias subjetivas definidas por el sistema causal que lo rige. Ante la posible réplica de que eso no puede ser considerado conciencia, afirma que si eso sucede, será debido a que la naturaleza de la experiencia del termostato es aún más lejana que la de un murciélago, y que será un prejuicio antropocéntrico el que no permita aceptar la experiencia subjetiva de un termostato. Por lo anterior, Chalmers considera que no hay razón fundamentada para dejar fuera del concepto de conciencia ese tipo de experiencia no biológica.

información: cuando el molde contiene un uno y cuando el molde contiene un cero.

En un proceso inverso, Chalmers propone reificar esos conceptos tomando en cuenta que los procesos en el mundo físico generan o se encuentran ligados a cierta información. En sus palabras (Chalmers, 1996, pp. 356):

> Los estados físicos corresponderán a los estados de información de acuerdo con sus efectos en el camino causal. Cuando dos estados físicos tienen el mismo efecto sobre el camino (causal) corresponderán al mismo estado de información. Si dividimos los estado físicos de este modo, llegaremos a un conjunto básico de diferencias físicas que hacen una diferencia[8]; esto constituye la *realización (realization)* física de la información.

Dada la concepción anterior, es posible afirmar que la información no se encuentra ligada a una única realización, sino que puede instanciarse de innumerables formas, siempre y cuando el espacio informacional corresponda a algún modelo causal en el mundo. Al extrapolar este argumento hacia las experiencias subjetivas, Chalmers afirma que podemos también encontrar información que se *realiza (realize)* en nuestra fenomenología, puesto que la misma información que se realiza en el mundo físico, también tiene un efecto causal en cada individuo al provocar la experiencia subjetiva asociada a la realización física.

Por otra parte, recordemos que el dualismo naturalista postula la imposibilidad de la explicación reductiva de la conciencia fenoménica, razón

8 Chalmers propone una reconocer a la información como *la diferencia que hace una diferencia*, refiriéndose a que, de una serie de estados (posiblemente de un continuo), la medición de un cambio de estado, relevante para determinado criterio, es lo que se convierte en información. Por ejemplo: un interruptor de luz tiene una serie de estado en un continuo (las posiciones entre encendido y apagado), pero solo es relevante cuando se genera un cambio en el estado del foco (encendido y apagado), por lo cual solo hay 2 estados relevantes y la información será encendido o apagado.

por la cual sería imposible que alguna realización física (como un estado funcional del sistema nervioso) pudiera considerarse como realización fenoménica. Entonces, la información que pueda ser realizable en el mundo físico y que tenga algún efecto en la conciencia fenoménica, constará por lo menos de dos realizaciones, es decir, esa información tendrá al menos dos aspectos.

Ante lo anterior, la apuesta de Chalmers apunta a buscar los mecanismos básicos que puedan enlazar a las partículas fundamentales de la conciencia con las partículas fundamentales de la física. Retomando el punto de vista informacional, Chalmers sugiere que la información puede jugar un papel principal en los mecanismos buscados. Ya en terreno completamente especulativo, Chalmers recuerda que existen teorías físicas como las postuladas por Wheeler (1994) y Fredkin (1990) que consideran a la información como ontológicamente fundamental para la física del universo, lo cual sería compatible con el dualismo naturalista.

Chalmers es explícito al mencionar que este último principio psicofísico es el más especulativo de todos y que no tiene argumentos contundentes para su demostración, por lo cual invita a considerarlo, más que como un argumento incompleto, como una estructura de la manera en que podrían construirse argumentos precisos. También menciona que si este principio pudiera ser comprobado de alguna manera, sería el más contundente y de mayor soporte a la teoría del dualismo naturalista, pues al describir los mecanismos de interacción de las partículas fundamentales de la conciencia con las partículas fundamentales de la física, la teoría fundamental estaría completa.

En el comentario anterior pareciera que Chalmers se contradice, postulando una interacción entre partículas de aspectos distintos, pues al interactuar, ambas pertenecen ya un mismo aspecto. Chalmers (1996, p. 381) comenta al respecto:

> La expresión doble aspecto debe interpretarse de un modo deflacionario: es meramente un modo colorido de hablar acerca de dos tipos diferentes de propiedades correlacionadas con una estructura similar. La información es simplemente una herramienta útil para caracterizar esta estructura común; no corresponde a una característica ontológica "profunda".

Lo cual es una manera de debilitar su versión del doble aspecto evitando la contradicción mencionada (aunque es posible que con ello esté asumiendo una explicación reductiva).

3.4 La conciencia a la luz del dualismo naturalista

Retomando las expectativas enunciadas en la sección 1.3, ofrezco a continuación una lectura de la forma en que el dualismo naturalista responde a sus preguntas asociadas:

(a) ¿qué es la conciencia?

La conciencia es el fenómeno implícito que experimenta individualmente cada sujeto o sistema al encontrarse en un estado funcional específico; en los seres humanos es especialmente relevante el estado de su sistema nervioso central. La conciencia se rige por una serie de principios psicofísicos, los cuales establecen la relación que tendrá aquella con el sustrato físico del sistema que lo contiene y con las funciones del mismo.

(b) ¿cómo es que existe?

La conciencia es una propiedad fundamental existente en donde haya interacción causal e información, por lo cual no tiene explicación en términos más básicos, aunque se busca una interacción legal entre esas partículas básicas y las de la física. Conforme aumenta el grado de complejidad de los sistemas, la conciencia en los componentes de los mismos toma formas igualmente complejas, generando así una experiencia más rica y vasta.

(c) ¿por qué y para qué existe la conciencia?

Chalmers no es explícito en la respuesta a esta pregunta, pues su enfoque especial es en las otras tres. Sin embargo, el apelativo de *naturalista* a su propuesta dualista nos proporciona una idea de la respuesta. Chalmers menciona que la conciencia superviene *naturalmente* de lo físico (aunque no superviene lógicamente), lo cual se puede interpretar como un compromiso con el punto de vista científico, donde la conciencia como la experimentamos es un *producto emergente* de la complejización de los sistemas que le dan sustrato, lo cual pareciera apelar de alguna manera a la evolución, si consideramos que ésta implica una complejización de los sistemas biológicos enfocada (entre otras cosas) a la adaptación de las especies para su supervivencia.

(d) ¿por qué y para qué el fenómeno de la conciencia lleva implícita una experiencia subjetiva ligada a cada experimentador individual?

Esta respuesta está íntimamente ligada a las dos primeras, puesto que la concepción de conciencia en esta teoría se encuentra enfocada a la experiencia subjetiva a la que refiere la primera pregunta, mientras que la segunda respuesta justifica la incapacidad de la explicación reduccionista, determinando que la conciencia es *per se* subjetiva.

Por último, propongo una forma de ubicar al dualismo naturalista en el panorama de las teorías de la conciencia como un dualismo de propiedades fundamentales que retoma elementos del funcionalismo para el abordaje científico de la conciencia a través de los principios psicofísicos que postula. Una interpretación de esto es afirmar que el funcionalismo es una parte indispensable del dualismo naturalista, o que el dualismo naturalista es una extensión del funcionalismo para poder explicar la conciencia fenoménica. Por otra parte, la idea de propiedades fundamentales hace que el dualismo naturalista comparta algunos puntos de vista y consecuencias con el panpsiquismo, pero con un enfoque cientificista. Se podría decir que el dualismo naturalista es una especie de *panpsiquismo funcionalista*.

4. La implementación de una computación.

Con los elementos principales del dualismo naturalista propuesto por Chalmers enumerados en el capítulo anterior, a continuación presento una la propuesta del mismo autor donde explica la forma en que dicha teoría de la conciencia es compatible y apoya las ambiciones de la IA fuerte, específicamente con respecto a la existencia de un tipo de cómputo suficiente para *ser* una mente con experiencias conscientes. Dicha propuesta inicialmente aparece publicada como un artículo llamado *On implementing a computation* (Chalmers, 1994), el cual después pasó a formar parte el capítulo 9 del libro *The Conscious Mind* (Chalmers, 1996) como una aplicación del dualismo naturalista.

Haciendo una recapitulación del dualismo naturalista, es posible observar que todos los postulados presentes en la teoría serían compatibles con la IA fuerte. En principio, la idea de las partículas fundamentales de la conciencia y el panpsiquismo resultante, dejan de lado la necesidad de algún tipo de mecanismo biológico necesario para la existencia de la conciencia en un agente. Si bien la coherencia estructural no tiene nada que afirmar al respecto, tampoco presenta alguna contradicción. Por su parte, el principio del doble aspecto de la información encuentra una gran compatibilidad en la IA fuerte, pues apunta a buscar en la información a las partículas fundamentales de la conciencia, y en la actualidad, el uso paradigmático de las computadoras es como mecanismos procesadores de información[1]. Por lo anterior, los sistemas computacionales serían agentes conscientes de inicio (por su naturaleza informacional) y la complejidad alcanzada por los mismos parecería potencialmente capaz de acercarse a la complejidad presente en la mente humana (dependiendo de la posición asumida en la IA fuerte).

Sin embargo, es en el principio de invariancia organizacional que dualismo naturalista encuentra el mayor apoyo con respecto a los objetivos de la IA fuerte, pues recordemos que dicho principio postula que dos

1 De allí que a la actividad relacionada con el uso de las computadoras como herramientas se le conozca con el nombre de *informática*.

sistemas con la misma organización funcional en detalle tendrán experiencias cualitativamente idénticas. Entonces, *si fuera* posible capturar la organización funcional del sistema cognitivo humano con el suficiente detalle y después replicarla en un sistema computacional, *entonces*, según el dualismo naturalista, el sistema computacional resultante tendría experiencias subjetivas idénticas a las de un ser humano. Mediante las cursivas anteriores pretendo resaltar que el argumento es un condicional extraído de los postulados de dualismo naturalista. Ya rebasada la parte correspondiente a la presentación y defensa del argumento teórico, Chalmers argumenta a favor de la existencia de la IA fuerte, dado el potencial soporte que le aportaría al dualismo naturalista. Para ello, la tarea de Chalmers consiste en defender el antecedente del condicional del párrafo anterior. Esto es que, para defender a la IA fuerte desde el ámbito del dualismo naturalista, Chalmers debe argumentar a favor de los siguientes puntos:

1.Es posible determinar el nivel de detalle del sistema cognitivo humano que sea capaz de dar cuenta de las experiencias subjetivas.

2.Es posible capturar en un modelo formal la organización funcional del sistema cognitivo humano con el suficiente detalle.

3.Es posible replicar la organización funcional del sistema cognitivo humano con el suficiente detalle en un sistema computacional,

los cuales presento y analizo en la siguientes secciones.

Llegados a este punto, creo necesario recordar que mi objetivo en esta tesis no es argumentar en contra de los postulados del dualismo naturalista, sino evidenciar los presupuestos y problemas subyacentes en la aplicación de la IA fuerte, por lo cual en el siguiente análisis tomo provisionalmente como correctos los postulados teóricos y me enfoco en la parte aplicativa.

4.1 El nivel de detalle necesario.

Previo a la exposición del mecanismo que propone Chalmers para capturar la organización funcional del sistema cognitivo humano, es necesario hacer explícito cuál es el nivel de detalle suficiente considerado por

el autor para poder dar cuenta de la mente humana. El objetivo de ello es, por una parte, evaluar la capacidad y el posible éxito de dicho mecanismo para lograr un modelo isomorfo causal de la mente humana y por otra parte, extraer los supuestos inmersos al postular ese nivel de detalle. Para ello, cito las palabras de Chalmers (1996, pp. 316):

> El tipo apropiado de organización funcional de un sistema estará siempre en un nivel lo suficientemente fino como para determinar sus capacidades conductuales. Llamemos a esta organización una organización funcional de *grano fino*. Para propósitos de ilustración, usualmente me concentraré en el nivel neuronal de organización en el cerebro, aunque podría bastar un nivel superior, y no es imposible que pudiese requerirse uno inferior. De cualquier manera los argumentos pueden generalizarse... Cuando dos sistemas comparten su organización funcional ..., diré que son isomorfos funcionales.

Como bien lo establece Chalmers, apelando a que los argumentos pueden generalizarse, no establece un compromiso con un nivel de detalle específico, y toma como referencia meramente ilustrativa el nivel de organización neuronal en el cerebro. Si por casualidad Chalmers hubiera acertado al afirmar que el nivel neuronal en el cerebro es suficiente para modelar un sistema isomorfo funcional, sería necesario crear un modelo formal del funcionamiento individual de cada neurona utilizada por el cerebro humano que juegue un papel relevante en la conducta, lo cual queda lejos del alcance de la neurociencia en la actualidad. Ahora que si el nivel de detalle necesita ser más profundo, digamos a nivel de partículas subatómicas, la factibilidad de la existencia de dicho modelo se vuelve cada vez menor. Si bien este no es un argumento contundente que niegue la imposibilidad del modelado formal de la causalidad mental, lo que quiero dejar claro es que existe la posibilidad de que no se pueda llegar a capturar dicho modelo causal, y que en esta propuesta, Chalmers requiere de dicho modelo, por lo

cual asume su existencia y es necesario que el lector haga lo mismo para seguir el argumento.

4.2 El modelo formal

Una vez que se ha asumido la existencia de un modelo formal de la cognición humana al nivel de detalle necesario, es necesario presentar el modelo lógico general que Chalmers propone como *molde* para establecer cualquier organización funcional capaz de dar cuenta de la causalidad del sistema cognitivo humano. El modelo en cuestión se presenta con el nombre de *Autómata de Estados Combinatorios* (AEC), el cual retoma el concepto computacional de *Autómata de Estados Finitos*[2] (AEF) y lo extiende para tener la capacidad de descomponer los estados presentes en un AEF en una serie de subestados, los cuales a su vez se pueden descomponer en más subestados, y así sucesivamente hasta alcanzar el nivel de detalle suficiente para capturar la *organización funcional* del sistema en cuestión, en este caso, el sistema causal subyacente en la mente humana que dé cuenta de la conducta. Es importante recalcar que la estructura lógica del AEC está diseñada específicamente para acoplarse a la definición de *organización funcional* definida en la sección 3.3.2.

De lo anterior sobresalen dos puntos: por un lado, al considerar que el funcionamiento de la mente se puede capturar mediante un AEC, Chalmers parece comprometerse a un modelo representacionalista de la mente, pues el mecanismo del AEF (del cual proviene el AEC) tiene su origen en la teoría de los lenguajes formales desarrollada principalmente por Chomsky y *Schützenberger* (1963), y dicha teoría se enfoca a analizar los lenguajes formales por sus características sintácticas, siendo tales símbolos *representaciones* arbitrarias de entidades con significado. A pesar de esto, Chalmers afirma que el modelo propuesto *es igualmente compatible con los*

2 Un AEF es una máquina abstracta compuesta por un número finito de estados, transiciones entre esos estados y acciones, la cual es un modelo abstracto de una máquina con memoria interna. En la jerarquía de gramáticas formales propuesta por Chomsky y *Schutzenberger* (1963), los AEF son máquinas de Turing que cumplen con ciertas restricciones para producir lenguajes con características muy precisas conocidos como *gramáticas regulares*.

enfoques simbólico y conexionista de la cognición, y también con otros enfoques computacionales, y lo anterior lo sostiene apelando a que *las concepciones computacionales pueden capturar casi cualquier tipo de organización causal* (Chalmers, 1996, pp. 332), pero no muestra en forma explícita la manera en que el conexionismo u otras posiciones filosóficas acerca de modelos de la mente pueden ser representadas mediante el AEC.

Una manera personal de interpretar la aseveración de Chalmers acerca de que las concepciones computacionales pueden capturar casi cualquier tipo de organización causal, es apelando a que, a pesar de las diferencias entre el representacionalismo, el conexionismo, la vida artificial y los sistemas basados en conducta, todos ellos utilizan como sustrato el hardware de arquitectura Von Neumann y para todos los casos hay resultados empíricos que permiten evaluar como exitosa la forma de capturar sus organizaciones causales. Sin embargo, al definir el funcionamiento del AEC, Chalmers adopta la concepción de conocimiento inmersa en el representacionalismo, y sin una explicación de la manera en que esa representación específica puede a su vez manipularse como red neuronal u otras estructuras subyacentes a las diversas concepciones de IA, no parece haber forma de dar cabida más que al representacionalismo.

El segundo punto a resaltar es la naturaleza del AEC como una máquina de Turing[3] (con ciertas características particulares). Lo anterior implica que el AEC también está sujeto al problema de la detención[4] y por tanto limitado a resolver problemas computables. Una vez mas, Chalmers está al tanto de la situación y es cauto al afirmar que la propuesta con respecto a su modelo debe leerse como un condicional: si la dinámica cognitiva es computable, entonces el tipo correcto de organización computacional dará origen a la conciencia. Dicho condicional muestra que

3 En realidad, lo que es una máquina de Turing por definición es el AEF. Pero al ser el AEC una versión del anterior, también es una máquina de Turing.

4 El problema de la detención para máquinas de Turing consiste en lo siguiente: dada una máquina de Turing M y una palabra w, determinar si M se detendrá cuando es ejecutada usando w como dato de entrada. Alan Turing (1936), demostró que el problema de la detención de la máquina de Turing es análogo al problema de la decisión.

esta propuesta depende fuertemente de la computabilidad de la mente, la cual Chalmers asume de forma implícita, pues no propone alternativa alguna a la posibilidad de que la mente no sea computable (por ejemplo, una propuesta posible es postular que la mente ocupe un grado mayor en la jerarquía aritmética de Kleene[5]), por lo cual, para seguir con el argumento, es necesario asumir la computabilidad de la dinámica cognitiva, pasando por alto objeciones como la que presenta Penrose (1989).

En resumen, para aceptar al AEC como el modelo general para capturar la causalidad del sistema cognitivo humano, es necesario aceptar las siguientes afirmaciones:

- La causalidad que rige el sistema cognitivo humano es computable.
- El sistema cognitivo humano usa representaciones simbólicas arbitrarias.

Dadas las características de las concepciones de la IA presentadas en el capítulo 2, parece haber una única opción para lo que se necesita asumir: el representacionalismo.

4.3 Replicar la organización funcional de la mente en un sistema computacional.

Una vez que asumida la posición representacionalista en el proyecto de IA fuerte defendido por Chalmers, se puede validar el tránsito de un modelo físico (en este caso el modelo de la causalidad en la cognición humana) a un modelo abstracto (el AEC resultante). Sin embargo, es necesario aún validar un último paso referente al tránsito de un modelo abstracto a su realización computacional.

Atinadamente, Chalmers enuncia que el principal problema que presenta la tarea de extraer conclusiones de los sistemas abstractos (como los sistemas matemáticos, programas de computadoras, máquinas de Turing, etcétera) y extrapolarlas a sistemas concretos (como los sistemas físicos y los cognitivos), es que se requiere un puente que relacione ambos dominios en

5 La jerarquía aritmética (Kozen, 1997) se refiere a una clasificación de las fórmulas en grados de insolubilidad, donde la idea es demostrar que los problemas no computables obedecen a una jerarquía. Una interpretación burda de ello es decir que, entre los problemas no computables (insolubles por una máquina de Turing), hay unos *más insolubles* que otros (Zenil, 2005).

esa dirección, es decir, que se requiere de un mecanismo que permita transitar de una manera adecuada del dominio abstracto hacia el concreto, para lo cual Chalmers postula un uso propio del concepto de implementación que se explica a continuación, para continuar comparándolo con el uso convencional de dicho concepto y los problemas que evidencian dichas diferencias.

4.3.1 El uso del concepto de implementación por parte de Chalmers.

Como se menciona arriba, Chalmers apuesta a que una concepción del uso del concepto de implementación funja como el puente necesario entre los dominios abstracto y concreto, para de esa manera validar la afirmación a cerca de que es posible replicar la organización funcional de la mente en un sistema computacional. Su propuesta es la siguiente (Chalmers, 1996):

> Un sistema físico *implementa* una computación cuando la estructura causal del sistema refleja la estructura formal de la computación. Es decir, el sistema *implementa* una computación si existe un modo de poner en correspondencia los estados del sistema con los estados de la computación, de manera que los estados físicos que estén causalmente relacionados se apliquen en estados formales que estén formalmente relacionados del modo correspondiente.

Si una vez más recordamos la definición básica de Chalmers de organización funcional (patrón abstracto de interacción causal), será posible ver la gran similitud que existe entre esta definición y la de implementación, siendo la diferencia más importante que el modelo abstracto en cuestión se encuentra realizado formalmente en una computación. Recordemos además que, asumido que Chalmers se apega al modelo de la IA fuerte especificado por el representacionalismo, es el AEC el modelo lógico propuesto para capturar el patrón causal de la estructura cognitiva humana. Al ser el AEC un modelo computable y definible mediante una máquina de Turing, es posible

afirmar que de hecho el AEC es ya un modelo computacional y que la manera de realizarlo computacionalmente solo es un trabajo de codificación en algún lenguaje de programación.

Al parecer, la propuesta de implementación de Chalmers no parece mas que manera abreviada de afirmar que el principio de invariancia organizacional del dualismo naturalista es directamente aplicable a la IA fuerte, siempre y cuando se acepte una concepción representacionalista de la misma, dada la computabilidad de la dinámica cognitiva y el uso de símbolos arbitrarios. Sin embargo, la adopción de esta manera de concebir la implementación es una redefinición *ad hoc* para apoyar al dualismo naturalista. Dado lo anterior, el siguiente paso para evaluar el concepto de implementación de Chalmers es compararlo con el uso de dicho concepto fuera del dualismo naturalista, para revisar las diferencias entre ellos y las posibles consecuencias y problemas que acarrearía la definición propuesta.

4.3.2 El uso convencional del concepto de implementación

La definición del diccionario del término *implementación*, apunta a designar alguna acción o los medios que permitan llevar a cabo o completar una tarea (Real Academia Española, 2001), y en el idioma inglés la referencia es análoga. Sin embargo es más común que el uso de dicho término se aplique a situaciones más específicas para designar el paso de un estado u objeto abstracto a uno más concreto o más detalladamente especificado. Así por ejemplo, la implementación de un procedimiento industrial se refiere a poner en práctica los mecanismos contenidos en una idea previamente establecida; la implementación de un algoritmo significa escribir el programa computacional que lleve a cabo los mecanismos específicos definidos en una descripción general. Al considerar que la palabra implementación es derivada del verbo transitivo *implementar*, la construcción gramatical de la forma *sujeto implementa objeto* se usa cuando el *sujeto* en cuestión es la entidad menos abstracta (más concreta o descrita en mayor detalle) resultado de las acciones o medios originados a partir del *objeto*. Una lectura de lo anterior es considerar que el *sujeto* responde a la descripción de una pregunta *¿cómo?*,

cuando el *¿qué?* se conoce previamente mediante el *objeto*. Es de esa manera que se entienden frases como las siguientes: *el navegador web implementa el protocolo http*, *la teoría de juegos implementa la lógica dinámica epistémica*, *la electrónica implementa conocimientos de la física*.

4.3.3 El problema del criterio de abstracción.

Lo expuesto anteriormente con respecto al uso convencional del término implementación parece en primera instancia concordar con la propuesta de Chalmers con respecto a la idea de que los objetos concretos implementan a los abstractos. En los usos habituales del término, subyace la idea no formalizada analizada arriba de que, cuando se dice que *x implementa y*, *x* será siempre más concreto que *y*. Sin embargo, si el concepto en cuestión ha de usarse para fundamentar una parte de una teoría filosófica relacionada con la IA, aparece la necesidad de establecer un criterio preciso y formal que permita diferenciar o medir cuantitativamente el grado de abstracción de los objetos, el cual no se encuentra especifica en la obra de Chalmers[6], dando origen al problema que nombraré en adelante como *el problema del criterio de abstracción*.

Debido al problema del criterio de abstracción, el uso del concepto de implementación entra en un terreno que da lugar a ambigüedades problemáticas para una teoría filosófica. Una de esas ambigüedades se presenta cuando existen situaciones donde, tanto el sujeto como el objeto (asociados al verbo implementar mencionado arriba) son entidades completamente abstractas. Un ejemplo de esto lo podemos encontrar el criterio para definir el grado de abstracción de un programa computacional. El uso común de implementación en computación apunta a que un programa computacional implementa cierto modelo lógico o matemático, dando por hecho que el programa es más concreto que el modelo de donde proviene.

6 Existen tratados al respecto, como el que presenta Edward Zalta (1983), donde se propone una teoría basada en la lógica modal para postular una metafísica que permite evitar postular una ontología especial asociada a los objetos abstractos diferente a la de los objetos concretos. Sin embargo, independientemente de la ontología de los objetos abstractos, la teoría de Chalmers necesita el establecimiento del criterio mencionado.

Pero ¿por qué deberíamos considerar a un programa computacional más concreto que otro modelo matemático y no simplemente una representación distinta del mismo modelo?

Para ejemplificar el problema del criterio de abstracción, propongo el caso de la teoría de la computabilidad esbozada en la sección 2.2: la máquina de Turing y el cálculo lambda son modelos matemáticos propuestos para abordar el problema de la decisión y fueron desarrollados de manera independiente basados en la misma idea, por lo cual ambos serían implementaciones con el mismo nivel jerárquico. Sin embargo, como se puede observar en la cita de Gandy de la misma sección, dicho autor (como gran parte de los teóricos de la computabilidad) considera que el cálculo lambda es más abstracto que la máquina de Turing, lo cual sería suficiente para definir que la máquina de Turing implementa el cálculo lambda. Para Sieg (2006), dicho criterio responde tanto a presupuestos filosóficos inmersos en los métodos matemáticos como a razones psicológicas implícitas en los individuos más que a cualquier criterio objetivo.

4.3.4 La causalidad como criterio de abstracción

En un intento de encontrar una manera en que la concepción de implementación de Chalmers pueda afrontar el problema del criterio de abstracción, es necesario buscar en su propuesta un criterio implícito, ya que no se indica explícitamente. Así, puede observarse Chalmers postula un elemento no presente en la concepción común de implementación: la causalidad. A continuación cito nuevamente un parte de la noción de implementación en Chalmers (p. 60): *el sistema implementa una computación... si los estados físicos que estén causalmente relacionados se aplican en estados formales que estén formalmente relacionados.*

Mediante la afirmación anterior, Chalmers parece asumir que una realización computacional es una representación formal de un sistema físico del cual se ha determinado capturado su causalidad, siendo ésta última el elemento que marca la diferencia entre lo abstracto y lo concreto. Recordando que el objetivo al momento es una búsqueda del criterio de

abstracción en la propuesta de Chalmers, podría considerarse que es la causalidad el elemento novedoso que aporta en su definición y que es la manera de reconocer los objetos abstractos de los concretos. Una lectura de lo anterior, es decir que Chalmers propone que los sistemas que son causalmente sensibles pertenecen al dominio de los objetos concretos, mientras que los sistemas que representen o simulen la causalidad mediante modelos formales, pertenecen al dominio de los objetos abstractos.

Si bien con lo anterior parece ser suficiente evadir el problema del criterio de abstracción, cuando se retoma dicha concepción en el uso común del término, trae como consecuencias diversas dificultades. Una de ellas es la manera de distinguir *grados* de abstracción, pues existen situaciones donde, tanto el objeto como el sujeto pertenecen al dominio de los entes abstractos, y aún así existe una intuición de que uno es más abstracto que el otro. Retomando el ejemplo de la máquina de Turing y el cálculo lambda: ninguno de esos sistemas es evidentemente más sensible a la causalidad que el otro, pero para los matemáticos es casi indiscutible que la máquina de Turing es el sistema más concreto. Ese mismo problema se presenta cuando se afirma que un algoritmo computacional es más concreto que un modelo matemático, por ejemplo, un programa de simulación del movimiento parabólico es tan sensible a la causalidad como la representación del mismo mediante operaciones geométricas y dibujos en papel. En todo caso, la noción de abstracción se debe referir a criterios de la metodología y la psicología de las personas expertas en el área, tal como lo plantea Sieg para la máquina de Turing y el cálculo lambda.

Entonces, cuando se utiliza el concepto de implementación y no existen una referencias explícitas a objetos del *mundo real* sensibles a la causalidad, no hay manera en que sea dicha causalidad la forma adecuada en distinguir entre los dominios abstracto y concreto. Extendiendo este argumento, es posible afirmar que una entidad abstracta puede definir un sinnúmero de implementaciones, por ejemplo la lógica dinámica epistémica puede ser implementada tanto por la teoría de juegos, como por algunos protocolos de comunicación en redes de computadora, por la programación

orientada a agentes, etcétera. El problema del criterio de abstracción emerge nuevamente cuando se necesita establecer una jerarquía entre los diversos sistemas realizados, es decir, cuando pretendemos compara el nivel de abstracción de la teoría de juegos con el nivel de abstracción de la programación orientada a agentes. Dado que ambos sistemas son insensibles a la causalidad, es evidente que ésta última no puede fungir con el rol de clasificación que pretende Chalmers en su propuesta.

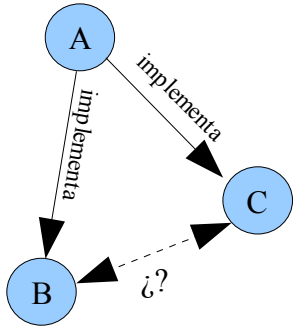


Ilustración 3. El problema del criterio de abstracción: Si B implementa A y C implementa A ¿Cómo medir la relación entre B y C?

4.4 La implementación de Chalmers

A lo largo de este capítulo, he intentado mostrar algunos de los presupuestos que necesita la aplicación de la teoría de Chalmers con respecto a la IA fuerte. Entre ellos resaltan:

- El ser humano puede conocer el nivel de detalle necesario para modelar el sistema cognitivo humano.
- El ser humano puede modelar formalmente el sistema cognitivo humano mediante un AEC.
- El modelo del sistema cognitivo humano es computable.
- La mente humana funciona como lo establece el representacionalismo.

Aceptando todo lo anterior, para que un sistema computacional tenga experiencias conscientes, es necesario también utilizar una noción de implementación donde la causalidad es un elemento crucial para relacionar al dicho sistema computacional con la mente humana. Sin embargo, dicha definición resulta problemática cuando se compara con los usos comunes del concepto, pues especialmente en el área de la computación, es difícil encontrar en todos los casos de implementaciones, referencias directas a sistemas sensibles a la causalidad.

Los problemas que presenta la definición propuesta por Chalmers, no permiten un uso generalizado de la misma, por lo cual sigue haciendo falta el puente que permita transitar del modelo abstracto hacia un modelo concreto, para lo cual es necesario comenzar con la definición de un criterio objetivo y deseablemente cuantitativo que permita ponderar los grados de abstracción de los objetos y definir de manera precisa una relación específica entre los ámbitos abstracto y concreto.

Sin una manera precisa de poder relacionar los estados formales con los causales, con los elementos existentes no hay manera de evaluar si un sistema computacional es isomorfo causal de un sistema cognitivo humano, por lo cual es imposible saber si ambos sistemas son invariantes organizacionales. En consecuencia, sin un criterio de abstracción preciso, aún aceptando los postulados del dualismo naturalista y los presupuestos con respecto a la IA enumerados arriba, no es posible afirmar que una computadora pueda tener experiencias conscientes idénticas a las de un ser humano.

5. Conclusiones y perspectivas

Hemos visto ya que el problema de la conciencia tiene una larga historia y se ha concebido de diversas maneras dependiendo del momento y el punto de vista que se asuman. Al día de hoy, el abordaje de dicho problema debe responder a preguntas desde diversas perspectivas, por lo cual toda teoría de la conciencia debe abarcar múltiples aspectos para poder ser considerada en los aspectos filosófico y científico. De entre todas las opciones plausibles, el dualismo naturalista propuesto por Chalmers es un candidato importante, pues intenta ocuparse del problema desde una perspectiva multidisciplinaria sin minimizar la importancia de los diversos aspectos del problema.

Después de un recorrido por algunas teorías de la conciencia, es posible ubicar al dualismo naturalista como una propuesta no reduccionista que permite mantener la perspectiva de la primera persona, apegándose a un criterio científico. Para lograrlo, postula una serie de principios psicofísicos (coherencia estructural, invariancia organizacional y doble aspecto de la información), apuntando a una manera de *atar* los datos objetivos que pueden ser generados por las disciplinas científicas con la experiencia subjetiva de cada individuo. Así, con la idea de extender y comprobar estos principios, Chalmers pretende ofrecer una base que pueda eventualmente convertirse en una teoría que de cuenta de la conciencia (especialmente lo referido a la experiencia subjetiva) a partir de datos objetivos provenientes de la ciencia empírica.

Como resultado de asumir los principios psicofísicos mencionados y la imposibilidad de la reducción, Chalmers postula una tesis ontológica donde se considera que la conciencia no puede ser explicada en términos más básicos, por lo cual debe considerarse como una entidad básica existente en los objetos del mundo. Lo anterior da como resultado que el dualismo naturalista sea compatible y asuma un panpsiquismo que puede encarnar en los sistemas computacionales.

Por otra parte, también hemos visto que desde la misma aparición de la computación, subyace una idea de capturar el mecanismo de cognición humana mediante un modelo formal, y que siendo la computadora digital el resultado de muchos años de desarrollo en ese sentido, se ofrece como el medio idóneo para poner a prueba la tesis de la IA fuerte. Sin embargo, al momento es imposible establecer un único absoluto modelo formal del pensamiento, pues así como las teorías de la conciencia, hay argumentos y datos empíricos que apoyan a los modelos computacionales del pensamiento como el representacionalismo, el conexionismo, la vida artificial simulada y los sistemas basados en conducta.

Con la idea de ofrecer un soporte desde diversos frentes al dualismo naturalista, Chalmers propone como una aplicación de su teoría la tesis de la IA fuerte. Sin embargo para ello asume, además del dualismo naturalista, una posición implícita con respecto al modelo computacional de la mente, que es básicamente lo que postula el representacionalismo.

Asumiendo el dualismo naturalista como teoría de la conciencia y el representacionalismo como concepción de la IA, Chalmers requiere también de una manera de relacionar al modelo formal de la IA con la estructura de la cognición humana. La idea es que mediante esa relación se pueda evaluar su isomorfismo funcional y apelar entonces al principio de la invariancia organizacional contenido en el dualismo naturalista, para afirmar que el dualismo naturalista da cuenta y es apoyado por la idea de la IA fuerte. Para ello, Chalmers propone una definición del concepto de implementación establecido con la finalidad de adecuarse a los conceptos del dualismo naturalista y al representacionalismo. Sin embargo, el uso de dicho concepto no captura la idea subyacente de lo que se quiere establecer mediante el uso cotidiano, específicamente en el área de la computación. El uso de la causalidad como criterio para reconocer y vincular a los sistemas abstractos de los concretos, es el punto donde el concepto de Chalmers ya no puede generalizarse, puesto que los sistemas computacionales no son necesariamente sensibles a la causalidad.

Si bien lo anterior muestra que la causalidad por sí sola no puede ser considerada con el criterio suficiente para vincular a los sistemas abstractos de los concretos, el problema es más general. La pregunta relevante no es si podemos apelar a la causalidad o no para clasificar a los objetos abstractos de los concretos, sino ¿a qué podemos apelar? Este cuestionamiento lo llamé el problema del criterio de abstracción.

Sin una definición objetiva y una medida cuantitativa precisa para medir el grado de abstracción de los objetos, es poco probable que llegue a ser útil alguna clasificación intuitiva. Pero aún más lejano se vé el objetivo de relacionar los objetos ya clasificados, especialmente si dicha relación implica una coherencia tan fuerte que pueda dar cuenta del principio de la invariancia organizacional. En consecuencia, aún aceptando los postulados del dualismo naturalista y los presupuestos con respecto a la IA enumerados arriba, no es posible afirmar que una computadora pueda tener experiencias conscientes idénticas a las de un ser humano, y por tanto no es posible afirmar que el dualismo naturalista de cuenta y apoye la existencia de la IA fuerte.

Hasta este punto, la argumentación ha apuntado a demostrar que la aplicación del dualismo naturalista con respecto a la IA fuerte no implica ningún tipo de soporte a dicha teoría. Sin embargo, las consecuencias parecen ir más allá. En el principio del doble aspecto de la información (la parte más especulativa del dualismo naturalista), Chalmers postula que la información puede tener innumerables realizaciones, entre ellas la realización fenoménica. Ya en la lectura más extrema, Chalmers invita a considerar que sea la información el punto ontológico donde convergen las realizaciones físicas con las fenoménicas y que es probable que sea la información donde exista ese puente tan buscado entre la conciencia fenoménica y la psicológica, apelando a teorías de la física que consideran también a la información como las partículas fundamentales.

Desde mi punto de vista, al llevar tan lejos su especulación, Chalmers se mueve libremente en un criterio de abstracción indefinido. A lo que me refiero es que, dado el problema del criterio de abstracción, Chalmers se permite reificar objetos abstractos (información) como partículas

fundamentales de la física y de esta manera *desordenar* una jerarquía ontológica que estaría bien establecida en presencia de un criterio de abstracción adecuado.

Con lo anterior quiero decir que comúnmente la manera en que se concibe la información desde el punto de vista humano, es como un proceso en el cual se extraen las características formales más relevantes de algún sistema (físico o menos abstracto) con la finalidad de recrear un modelo más cognoscible o manipulable. Entonces, si fuéramos capaces de establecer objetivamente la manera en que se relacionan ontológicamente un sistema físico con su producto abstracto, ese tipo de especulaciones no tendrían lugar.

Ampliando aun más el alcance del problema del criterio de abstracción con respecto al problema de la conciencia, es posible ver que afecta a cualquier teoría que haga uso de modelos abstractos, tales como la teoría de la identidad y el funcionalismo. En otras palabras, cualquier teoría que necesite fundamentar el puente entre los modelos abstractos y el funcionamiento físico de la mente humana, eventualmente llegará a la necesidad de establecer un criterio de abstracción como el que he mencionado, ya sea antes o mediante la definición del concepto de implementación.

Una vez esclarecida la importancia de lo anterior, espero que este trabajo logre mostrar la relevancia del problema del criterio de abstracción y de la manera en que debe ser concebido el concepto de implementación, para de esa manera poder dar pie a una solución más general mediante un criterio de abstracción más desarrollado. Probablemente, esta solución incluya a la causalidad como un parámetro necesario en el criterio, mas ya hemos visto que no es suficiente.

Con lo anterior en mente, una posible manera de continuar el proyecto esbozado en esta tesis, consistirá en realizar un análisis exhaustivo de las diferencias entre lo que se desea concebir como abstracto y concreto, para partir de allí a la construcción de un método cuantitativo que permita la clasificación y jerarquización de los objetos según su grado de abstracción. Pero más importante aun será apuntar hacia la búsqueda de una relación

ontológica entre los modelos abstractos y los sistemas físicos que representan, pues es ese punto donde las teorías de la conciencia encuentran un hueco explicativo difícil de superar. Para ello, será muy útil tomar en cuenta las consideraciones existentes hasta el momento en las teorías de la conciencia (incluyendo al dualismo naturalista), los tratados de los objetos abstractos como el de Zalta (1983) y de la naturaleza de los modelos en la ciencia como el de Díaz (2007, pp. 395-411).

Bibliografía

Boden, Margaret A., 1988, *Computer Models of Mind,* Cambridge University Press, Cambridge.

Boden, Margaret A. (ed.), 1989, *The Philosophy of Artificial Intelligence,* Cambridge University Press, Cambridge.

Boyd, R., 1980, "Materialism without reductionism: What physicalism does not entail." en Ned Block, ed. *Readings in the Philosophy of Psychology*, Vol. 1, Harvard University Press, Cambridge .

Brentano, Franz C., 1874/1973, *Psychology from an Empirical Standpoint*, Routeledge, Londres.

Brooks, Rodney, 1986, "Achieving Artificial Intelligence Through Building Robots", *MIT AI Lab Memo*, Num. 899, Cambridge.

Brooks, Rodney, 1991, "Intelligence without representation", *Artificial Intelligence Journal*, Num. 47, pp. 139-159.

Bunge, Mario, 2003, *Emergence and Convergence*, University of Toronto Press, Toronto.

Chalmers, David J., 1994, "On implementing a computation", *Minds and Machines*, num. 4, pp. 391-402.

Chalmers, David J., 1995/1997, "Facing up to the problem of consciousness", en Shear 1997, pp. 9-30.

Chalmers, David J., 1996, *The Conscious Mind: in search of a fundamental theory*, Oxford University Press, Nueva York. [Versón en castellano: 2003, *La mente consciente: en búsqueda de una teoría fundamental*, trad. José Diez, Gedisa, Barcelona]

Chomsky, Noam y *Schützenberger, Marcel P.*, 1963, *"The algebraic theory of context free languages", en* Braffort y Hirschberg (eds.), *Computer Programming and Formal Languages, North-Holland, Amsterdam, pp. 118-161.*

Churchland, Paul M., 1989, *A neurocomputational perspective: the nature of mind and the structure of science,* MIT Press, Cambridge.

Churchland, Patricia Smith, 1983, "Consciousness: the transmutation of a concept". *Pacific Philosophical Quarterly*, num.64, Los Angeles. pp. 80-95.

Clark, Andy, 1996, "Philosophical foundations", en Boden (ed.), *Artificial Intelligence*, Academic Press, San Diego.

Cole, David, 2008, "The Chinese Room Argument", *The Stanford Encyclopedia of Philosophy*, Zalta (ed.),

URL=<http://plato.stanford.edu/archives/spr2008/entries/chinese-room/>.

Davidson, Donald, 1995, *Ensayos sobre acciones y sucesos*, Crítica / UNAM-IIF, Barcelona.

Davis, Martin, 1987/1995, "Mathematical Logic and the Origin of Modern Computers", en Herken 1995, pp. 135-158.

Dennett, Daniel C., 1991, *Consciousness Explained*, Boston, Little, Brown and Company, Boston..

Descartes, René, 1637/1987, *Meditaciones Metafísicas y otros*, trad. E. López y M. Grana, Gredos, Madrid.

Díaz, José Luis, 2007, *La Conciencia Viviente*, Fondo de Cultura Económica, México.

Ezcurdia, Maite y Hansberg, Olbeth (comps.), 2003, *La naturaleza de la experiencia*, vol. 1, UNAM-IIF, México.

Fodor, Jerry A.,1983, *The modularity of mind :an essay of faculty psychology*, MIT Press, Cambridge.

Fodor, Jerry A. y Pylyshyn, Zenon, 1988, "Connectionism and cognitive architecture", *Cognition*, vol. 28, pp. 3-71.

Foucault, Michael, 1997, *Las Palabras y las Cosas: una Arqueología de las Ciencias Humanas*, Siglo XXI, Madrid.

Fredkin, E., 1990, "Digital Mechanics" en *Physica D*, vol. 45, pp. 254 – 270 [Citado en Chalmers 1996].

Gandy, Robin, 1995, “The confluence of ideas in 1936”, en Herken 1995, pp. 51-102.

Gardner, Howard, 1985, *The Mind's New Science: A History of the Cognitive Revolution*, Basic Books, Nueva York.

Gödel, Kurt, 1929/1990, “On the completeness of the calculus of logic”, en Feferman et al. (eds), *Collected Works of Kurt Gödel*, vol. 1. Oxford University Press, Oxford. pp. 61-101.

Gödel, Kurt, 1949/1990, “A Remark about the Relationship between Relativity Theory and Idealistic Philosophy”, en Feferman et al. (eds), *Collected Works of Kurt Gödel*, vol. 2. Oxford University Press, Oxford. pp. 202-207

Herken, Rolf (ed.), 1995, *The Universal Turing Machine A Half Century Survey*, 2a ed. Springer-Verlag, Viena.

Hernández, Victor M., 1999, “Leibniz y la lingua Characterica ”, *Diánoia*, vol 44, num 45, UNAM-IFF, México.

Hernández-Quiroz, Francisco y Morado, Raymundo, 2006, “Hilbert, Turing y la noción de procedimento efectivo”, *Ludus Vitalis*, vol. 14, num. 26, México.

Hodgson, David, 1997, “The easy problems ain't so easy”, en Shear 1997, pp. 125-131.

Husserl, Edmund, 1913/1949, *Ideas relativas a una fenomenologia pura y una filosofia fenomenológica*, trad. José Gaos, Fondo de Cultura Económica, México.

Jackson, Frank, 1982/2003, “Qualia Epifenoménicos”, trad. Laura E. Manriquez, en Ezcurdia y Hansberg 2003, pp 95-110.

Jamshidi, Mohammad y Zilouchian, Ali, 2001, *Intelligent Control Systems Using Soft Computing Methodologies*, CRC Press, Boca Raton, Florida.

James, William, 1890/1989, *Principios de Psicología*, trad. Agustín Bárcena, Fondo de Cultura Económica, México.

Jordan, Michael I. y Russell, Stuart, 1999, “Computational Intelligence”, en Wilson y Keil (eds.) *The MIT Encyclopedia of the Cognitive Sciences,* MIT Press, Cambridge.

Kim, Jaewong, 1998, *Mind in a Physical World*, MIT Press, Cambridge.

Kozen, Dexter C., 1997, *Automata and Computability*, Springer, Nueva York.

La Mettrie, Julien Offray de, 1748/1962, *El hombre máquina,* Eudeba, Buenos Aires.

Langton, Christopher, 1995,”Editor's introduction”, en *Artificial Life: an overview*, MIT Press, Massachusets, pp. ix -xi.

Lewis, David, 1972, “Psychophysical and theoretical identifications”, *Australasian Journal of Philosophy*, 50, 249-58.

Lowe, E.J., 1997,“There are no easy probles in consciousness”, en Shear 1997, pp. 117-124.

Luria, Alexander R., 1978, *Cerebro y lenguaje :La Afasia Traumática*, trad. Luis Flaquer, Fontanella, Barcelona.

Lutz, Antoine y Thompson, Evan, 2003, “Neurophenomenology: Integrating subjective experience and brain dynamics in the neuroscience of consciousness”, *Journal of Consciousness Studies* ,10, pp. 31-52.

Marr, David, 1982, *Vision,* Henry Holt and co., Nueva York.

Mathers, Carola B., 1986, “Psychoanalisis: Science or Nonscience?”, *Bulletin of the Royal College of Psychiatrists*, Londres, 10, pp. 103-104.

McCulloch, Warren S. y Pitts, Walter H., 1943, “A logical calculus of the ideas immanent in nervous activity”, en Boden 1989, pp. 22-39.

Merleau-Ponty, Maurice, 1942/1957, *La Estructura del Comportamiento*, trad. Enrique Alonso, Hachette, Buenos Aires.

Merleau-Ponty, Maurice, 1945/1957, *Fenomenología de la Percepción*, trad. Emilio Uranga, Fondo de Cultura Económica, México.

Nagel, Thomas, 1974/2003, “¿Cómo es ser un murciélago?”, trad. Héctor Islas, en Ezcurdia y Hansberg 2003, pp. 45-63.

Penrose, Roger, 1989, *The Emperor's new mind*, Oxford University Press, Nueva York.

Preston, J y Bishop, M., 2002, *Views into the Chinese room: new essays on Searle and artificial intelligence,* Oxford University Press, Clarendon.

Putnam, Hilary, 1967/1981, "La vida mental de algunas máquinas", trad. Martha Gorostiza, *Cuadernos de Crítica*, num. 17, UNAM-IFF, México.

Real Academia Española, 2001, Diccionario de la lengua española. 22a ed., Espasa Calpe, Madrid.

Robinson, Howard, 2003, 'Dualism', en Stich and Warfield (eds), *The Blackwell Guide to Philosophy of Mind*, Blackwell, Oxford, pp. 85-101.

Rumelhart, D y McClelland, J, 1987a, *Parallel distributed processing vol. 1: Foundations*, MIT Press, Cambridge.

Rumelhart, D y McClelland, J, 1987b, *Parallel distributed processing vol. 2: Psychologial and Biological Models*, MIT Press, Cambridge.

Rumelhart, D y McClelland, J, 1987c, "On learning past tenses in English verbs", en Rumelhart, D y McClelland, J, 1987b, pp. 216-271.

Sartre, Jean Paul, 1982, *Psicología Fenomenológica de la Imaginación*, trad. Manuel Lamana, Losada, Buenos Aires.

Shear, Jonathan (ed), 1997, *Explaining Consciousness: the hard problem*, Bradford Books / MIT Press, Cambridge.

Searle, John R., 1980/1989, "Minds, brains and programs", en Boden 1989, pp.67-88.

Sieg, Wilfred, 2006, "Gödel on computability", *Philosophia Mathematica*, num 14, pp. 189-207.

Sieg, Wilfred, 2008, "Church without dogma: axioms for computability", en Lowe et al. (ed). *New Computational Paradigms*, Springer Verlag, pp. 139-152.

Turing, Alan, 1950/1989, "Computing machinery and intelligence", en Boden 1989, pp. 40-66.

Sperry, R. 1977, “Forebrain commissurotomy and conscious awareness”, *Journal of Medicine and Philosophy*, 2, pp. 101-126 [Citado en Díaz 2007].

Uspenski, V. A., 1979/1983, *Máquina de Post*, trad. Stanislav N. Belousov, Editorial Mir, Moscú.

Varela, Francisco J., Thompson, Evan y Rosch, Eleanor, 1992, *The Embodied Mind: Cognitive Science and Human Experience*, MIT Press, Cambridge. [Citas y páginas de la versión en castellano: 2005, *De cuerpo presente: las ciencias cognitivas y la experiencia humana*, trad. Carlos Gardini, Gedisa, Barcelona,]

Von Neumann, John, 1958, *The Computer and the Brain,* Yale University Press, New Haven.

Wheeler, J.A., 1994, “It from bit” en *At home in the universe*, American Institute of Physics Press, Nueva York. [Citado en Chalmers 1996]

Winston, Patrick H., 1992/1994, *Inteligencia Artificial*, 3a ed., trad. Homero Flores Samaniego, Addison-Wesley Iberoamericana, Wilmington.

Wolfram, Stephen, 1994, *Cellular Automata and Complexity*, Addison-Wesley.

Woodward, William, 1972, “Fechner's panpsychism: A scientific solution to the mind-body problem” en *Journal of the History of Behavioral Sciences*, vol. 8, pp. 367-386.

Zalta, Edward, 1983, *Abstract Objects: An Introduction to Axiomatic Metaphysics*, D. Reidel.

Zenil, Héctor, 2005, *Encaje de las Redes Neuronales Recurrentes Analógicas en la Jerarquía Aritmética*. Tesis de la Facultad de Ciencias de la UNAM, México.

Zilouchian, Ali, 2001, “Fundamentals of Neural Networks”, en Jamshidi y Zilouchian 2001.