UNIVERSIDAD DE CHILE
Facultad de Filosofía y Humanidades
Escuela de Postgrado
Centro de Estudios Cognitivos

Conciencia e Inteligencia Artificial:

Consideraciones Críticas sobre la Plausibilidad de que una Máquina Programada Posea Conciencia Fenoménica

Tesis para optar al grado de Magíster en Estudios Cognitivos

Autor:
Cristóbal Fuentes Barassi
Profesor Patrocinante: Rodrigo González Fernández

Santiago de Chile 2011

Agradecimientos

A Elías por iluminar inmensamente mis días y ser el niño más bello del mundo, que junto a su alegría y compañía hace que todo tenga más sentido. A los profesores que fueron capaces de mirar a la persona detrás del alumno. Finalmente, a Dios por mostrar los caminos en los momentos correctos.

Resumen

El presente trabajo tiene como objetivo realizar observaciones críticas a la plausibilidad de que una máquina programada en función de algoritmos pueda poseer conciencia fenoménica.

Las observaciones críticas se vinculan con la implausibilidad de que la utilización de algoritmos sea suficiente para explicar y entregar experiencia subjetiva a una máquina programada. Entre las críticas se encuentra el hecho de que la experiencia cualitativa del mundo ocurre desde un punto de vista. Además, no existe - usando la terminología de Nagel (1974) - algo así como “la experiencia subjetiva de ser como una máquina”. También se explora cómo el carácter subjetivo de la conciencia fenoménica presenta dificultades adicionales para su modelamiento e implementación por parte de la Inteligencia Artificial, dada su ontología de primera persona. Finalmente, se entregan argumentos a favor de la importancia de entregar conciencia fenoménica a un artefacto artificial toda vez que ésta formaría parte esencial para la conducta inteligente, pero dicha tarea debería llevarse a cabo a través de otros modelamientos alternativos al uso de algoritmos, toda vez que éstos no son capaces de explicar la existencia de la conciencia fenoménica.

Epígrafe

***Professor Hobby** : [after stabbing the robot's hand in a demonstration to an audience about how real the robot was] How did that make you feel? Angry? Shocked? **Secretary** : I don't understand. **Professor Hobby** : What did I do to your feelings? **Secretary** : You did it to my hand. **Professor Hobby** : Ok... now tell me, what is love? **Secretary** : Love is first widening my eyes a little bit and quickening my breathing a little and warming my skin and touching... **Professor Hobby** : ...and so on. Exactly so. Thank you, Sheila.*
(De la película 'AI: Artificial Intelligence', 2001)

Introducción

¿Es posible crear una máquina programada que posea conciencia fenoménica? De manera más concreta, ¿Es plausible que un artefacto creado y programado en función de algoritmos por el hombre pueda tener experiencias subjetivas de color, dolor, aroma, disfrutar de una melodía o sentir estrés?

Desde el surgimiento de la Inteligencia Artificial (I.A.) con el trabajo seminal de Alan Turing (1948, 1950), la disciplina ha intentado crear máquinas capaces de emular y competir con las capacidades cognitivas del ser humano, utilizando como piedra angular el concepto de 'Algoritmo'. Un algoritmo, que definido *grosso modo* es una sucesión ordenada de pasos finitos que llevan a un resultado, es lo que yace detrás de lo que se denomina una 'Máquina de Turing', entendida como una máquina ficticia que posee diferentes estados, y dado determinado *input*, es capaz de alcanzar el próximo estado. La idea básica de esta aproximación clásica en I.A. es que si una máquina es dotada con el algoritmo apropiado, debiera ser capaz de realizar cualquier tarea que un humano puede hacer.

Es debido a este "optimismo" que ha habido un interés en la disciplina por crear máquinas capaces de imitar otras características del pensamiento humano, con especial atención al desarrollo de máquinas que exhiban conciencia (Gámez, 2007). Precisamente por eso, en la última década ha comenzado a surgir un creciente interés en lo que se ha denominado 'Conciencia Artificial' (también conocida como 'Conciencia de Máquina' y 'Conciencia Sintética'), la cual podría ser definida como el área de la I.A. que busca diseñar e implementar modelos de conciencia artificial. Las razones para tal empresa podrían agruparse en torno a cuatro motivaciones centrales: Primero, y en directa relación con el trabajo en Ciencia Cognitiva, diseñar modelamientos de la conciencia para determinar cuáles serían sus rasgos determinantes; segundo, que las implementaciones de dichos modelos pudieran servir como herramientas para probar y entender la conciencia humana, tercero, determinar cómo y cuál es la relación entre la conciencia y el soporte físico sobre la cual se produce, y finalmente, que la conciencia fenoménica forma parte fundamental en algunas de las tareas consideradas inteligentes[1].

Las artes han dado cuenta del "optimismo" antes mencionado, especialmente la literatura y el cine, y en sus obras han presentado máquinas altamente sofisticadas, capaces de realizar tareas complejas tales como manejar naves espaciales, jugar ajedrez, generar estrategias de guerra, identificar emociones y entregar juicios estéticos, siendo el mejor ejemplo *HAL 9000* en '2001: Odisea en el Espacio' de Stanley Kubrick, o quizás una versión algo más actual y beligerante en 'The Matrix' de los hermanos Wachowsky, o 'Terminator' de James Cameron. También se pueden señalar ejemplos menos violentos, pero que exhiben la misma idea, tales como 'I, Robot' de Isaac Asimov, libro que trata sobre un robot que desarrolla conciencia debido a un generador de *'self'* que le permite estar consciente, o 'AI: Artificial Intelligence' de Steven Spielberg, película que da cuenta de un futuro en el cual existen máquinas idénticas a los humanos que son conscientes de sí mismas y tienen la habilidad de aprender, tener sentimientos, y generar afectos hacia sus dueños. Al parecer, no cuesta tanto trabajo aceptar desde nuestra intuición las suposiciones que yacen detrás de estas obras de arte: la conciencia es algo que se puede entregar a

[1] Más sobre estas motivaciones en el capitulo 2.

una máquina al incluir una pieza o parte adecuada, o a lo sumo, es una característica que "emerge" al existir una semejanza suficiente a la configuración física de los humanos.

Aquello que resulta tan fácil de asimilar desde una *folk psychology*, aún es motivo de discusión en I.A. Esta disciplina aún no ha podido dar cuenta de ciertas cuestiones filosóficas relacionadas con la creación de máquinas que simulen características humanas, y la discusión se vuelve particularmente controvertida al plantear seriamente la admisibilidad de que exhiban conciencia fenoménica, entendida ésta como la capacidad de tener experiencias subjetivas de manera similar a como un ser humano las posee, tales como saborear un té, disfrutar de la música o sentir asco ante un olor nauseabundo. De lo anterior, se desprenden otras preguntas igualmente relevantes con respecto a problemáticas tales como cómo dilucidar si es que poseen dicha conciencia, cuál es la relación entre el soporte físico y la aparición de ésta, o si es que una máquina puede pensar o sólo realiza cálculos.

La I.A. sostiene que muchos de los avances en la disciplina aseguran la eventual creación de máquinas con capacidades cognitivas similares, o incluso mejores que las humanas (Alexander, 2007). Sin ir más lejos, el año 1997 se produjo un gran revuelo cuando *Deep Blue*, la máquina creada por IBM para jugar ajedrez, venció al entonces campeón mundial vigente Garry Kasparov. Por primera vez, la Inteligencia Artificial derrotaba a la inteligencia natural. Sin embargo, pareciera controvertido afirmar que *Deep Blue* entiende ajedrez, aunque es obvio que juega incluso mejor que un profesional;ésta simplemente aplica determinadas reglas para encontrar una jugada que lleve a una mejor posición de las piezas en el tablero, de acuerdo con un criterio de evaluación programado por jugadores profesionales de ajedrez, siguiendo los algoritmos entregados por los ingenieros y realizando una astronómica cantidad de cálculos. En otras palabras, *Deep Blue* entiende tanto lo que significa jugar ajedrez como un televisor podría entender las imágenes que se proyectan en su pantalla, ya que lo que la máquina hace realmente es seguir una serie de pasos previamente establecidos para luego ponerlos a funcionar[2]. *Deep Blue* no posee la experiencia subjetiva de jugar ajedrez que su contendor tiene, ya que nada en el diseño y funcionamiento de los algoritmos dota a la máquina de la experiencia subjetiva de jugar ajedrez, ni de otras experiencias asociadas a dicho deporte, tales como la sensación de tensión, de apuro, de sentir presión al estar perdiendo y las experiencias asociadas al observar al contrincante, experiencias que, a su vez, podrían modificar los patrones de juego de un ajedrecista. Defensores de lo que Searle (1980) denomina 'Inteligencia Artificial Fuerte' o visión más extrema de la I.A., sostendrían que sólo se necesita un parámetro de observación externa de conducta para determinar si una máquina presenta características tales como inteligencia u otros rasgos de la actividad inteligente que pudieran ser asociadas a ésta (de manera similar a lo planteado por Turing en el 'Test de Turing'); es decir, y ejemplificando lo anterior, una máquina sería consciente si es capaz de pasar las pruebas que nosotros, los seres humanos, determinemos como condiciones suficientes para establecer que dicha máquina presenta conciencia. Sin ir más lejos, añadirían, eso es lo que hacemos todo el tiempo con otros seres humanos, al asumir que ellos también tienen conciencia debido a que se comportan y lucen nosotros. Pero claramente, dicha afirmación - de un claro sesgo conductista - no puede emitirse con respecto a una criatura hecha de cables y silicio. La conciencia no es algo a lo que se acceda desde la objetividad de una prueba, y sólo el poseedor de esta conciencia puede establecer su presencia o ausencia. Esto último no sólo es un problema para la I.A., sino también para la Filosofía de la Mente, y es por esto, que en Filosofía de la Mente existe una variada gama de planteamientos sobre

[2] Más sobre esto en Apéndice.

cómo dar cuenta de este fenómeno. Dichas aproximaciones provienen de discusiones sobre el problema mente-cuerpo, cuyas principales escuelas de pensamiento son el Dualismo (mente y cuerpo son dos sustancias ontológicamente distintas) y el Monismo (sólo lo físico es real, y lo mental puede ser reducido a lo físico, o bien sólo lo mental es real). Si bien esta discusión recobra fuerza con Descartes a mediados del siglo XVII, en la actualidad aún no hay consenso, y muy por el contrario, se ha producido una ramificación de sub-argumentos derivados de los ya mencionados, contribuyendo aún más a una falta de unanimidad con respecto al tema. En relación con esto último, el reconocido neurobiólogo Bernard Baars resume esta confusión con las siguientes palabras:

Si hacemos preguntas sobre la conciencia puramente en términos de subjetividad - ¿qué se siente ser tú o yo? - uno se encuentra con la clásica paradoja mente-cuerpo en la cual se termina con una de las tres posiciones clásicas del problema: Mentalismo, Fisicismo o Dualismo; y el diálogo sobre estos tópicos - mejor dicho, el diálogo entre sordos - se da vueltas y vueltas y nunca se soluciona. Por lo tanto, desde mi punto de vista, la primera cosa que se debe hacer si es que de hecho se quieren responder preguntas, es plantearlas de una manera que se puedan responder. (En Blackmore, 2005, p.11, traducción mía.)

Es importante señalar que dentro de cada una de estas aproximaciones mencionadas por Baars, existen diversas posiciones, enfatizando diversos contenidos u otorgando preponderancia a distintas pruebas y argumentos, que en el Capítulo 1.2 serán más desarrolladas.

De manera adicional, se puede hacer alusión a otra dimensión del problema, que tiene que ver con una definición clara sobre qué hablamos cuando hablamos de conciencia. Ned Block (1995b) lo denomina un "Concepto Híbrido", ya que éste connota una serie de conceptos distintos y denota un número de fenómenos diferentes. Como Block señala, muchos filósofos, autores y disciplinas declaran estar tratando con el tema de la conciencia, cuando en realidad están tratando con una variedad de fenómenos relacionados a la experiencia subjetiva el mundo, tales como la distinción conciencia/inconciencia que indica la oposición entre estar despierto y no estarlo (ejemplo: "Luego del accidente el chofer quedó inconsciente"), o considerada como 'atención' a algo (ejemplo: "Estoy consciente de que el semáforo está en rojo") que al mismo tiempo puede ser conciencia de algo externo o de algo interno (ejemplo: "Estoy consciente de mis ideas sobre política"), o considerada como la habilidad de identificarse como individuo o *self-consciousness,* o como la habilidad de experimentar el mundo desde un punto de vista subjetivo (ejemplo: Sensaciones, *qualia* o manera en la que se sienten las experiencias tales como color, dolor), etc.

El presente trabajo realiza un análisis crítico sobre la plausibilidad de que una máquina programada en función de algoritmos posea conciencia fenoménica, lo que Nagel (1974) denomina *'what it is like to be'*, o la experiencia subjetiva del mundo que un ser experimenta al encontrarse en contacto con su medio ambiente y su vida interior. Es decir, para propósitos del presente trabajo, es importante tener en cuenta que otras nociones de conciencia pueden ser mencionadas, pero el análisis crítico que se presenta se centra específicamente en la plausibilidad de que una máquina programada exhiba conciencia fenoménica, sin ahondar en el hecho posible de que ésta pueda o no exhibir otros tipos de conciencia (tales como *self- consciousness*, Conciencia de Acceso u otras)[3]. Lo anterior resulta filosóficamente relevante ya que podría ayudar a una mejor comprensión de qué

[3] Más sobre distintos tipos de conciencia en capítulo 3.

es lo que hablamos cuando hablamos de conciencia, a analizar sus componentes claves, y una vez que se haga eso, discutir el eventual planteamiento de un modelamiento de la "experiencia subjetiva del mundo". Si es que tal cosa es posible, primero que todo se debe comenzar con la discusión de los fundamentos filosóficos que dan pie al trabajo de disciplinas tales como la I.A. o la Psicología. De manera adicional, la discusión sobre la conciencia fenoménica en máquinas programadas permite cuestionar cuáles son los límites éticos para embarcarse en tal empresa, considerando las ramificaciones legales y sociales que la creación de máquinas conscientes podría traer, las cuales no se analizarán en detalle en el presente trabajo dada la profundidad del tema, dejando abierta la posibilidad de que puedan ser exploradas en futuras investigaciones. Finalmente, se discute la importancia de dotar de conciencia fenoménica a máquinas programadas enfatizando el hecho de que ésta sí juega un rol fundamental en determinadas tareas consideradas inteligentes, a diferencia de planteamientos que la consideran sólo como un epifenómeno[4], irrelevante para el resto del sistema cognitivo. Es importante destacar que se sostiene una posición favorable al hecho que tal tarea es posible y necesaria. Sin embargo, se sostiene que el modelamiento a través de algoritmos pudiera no ser la mejor alternativa al momento de intentarlo, toda vez que el uso de algoritmos no da cuenta de cómo y por qué se produce este particular rasgo de la actividad cognitiva[5].

[4] Más sobre el concepto de epifenómeno en 1.2.2. Brevemente, aseverar que la conciencia es un epifenómeno sería plantear que si bien ésta es producida por el funcionamiento del cerebro, es causalmente irrelevante para el resto del sistema cognitivo.

[5] Más sobre este último punto en Capítulo 4.

1. Origen: El Problema Mente-Cuerpo.

1.1 El Lector en el Parque.

Para plantear y discutir las dificultades que la conciencia presenta a la Inteligencia Artificial, es necesario comprender el origen de estas dificultades, y éste es el denominado 'Problema Mente-Cuerpo'.

Para introducir este último, y en definitiva, el problema de la conciencia, apelaremos a un ejemplo cotidiano de experiencia personal. Suponga que Ud. se encuentra sentado en un parque, con un libro bajo el brazo, pensando sobre la economía nacional de los últimos meses. Luego, abre el libro que con anterioridad decidiera sacar de su biblioteca en la oficina, y procede a leer hasta la página 10. Su lectura se ve interrumpida, ya que un zancudo se ha posado sobre su mano, y le ha inyectado su ponzoñoso veneno. Siente una pequeña sensación de pinchazo en su dedo anular, y por tanto decide tomar el libro por la cubierta para aplastar al molesto insecto. El pinchazo que siente posee ciertas características, entre las cuales se encuentran el hecho de que sólo existe en la medida en que es conscientemente vivido por Ud., por lo que se podría decir que es un hecho completamente "subjetivo" (y no "objetivo"), ya que sólo Ud. y nadie más puede saber y experimentar la sensación que tiene en ese momento, y ninguna otra persona ni entidad desde el exterior puede experimentar la sensación en cuestión. También se podría mencionar una segunda característica de este dolor, y es el hecho de que esta sensación tiene un rasgo cualitativo, en este caso un rasgo cualitativo asociado al dolor, o si se quiere, a una molestia física. En resumen, se podría aseverar que el dolor asociado al pinchazo del zancudo tiene las características de ser *subjetivo* y de ser *cualitativo*.

Ahora bien, traslademos el foco de atención y consideremos los elementos que se encuentran en el parque. Los árboles, la banca sobre la cual está sentado y el libro en su mano. Todos estos objetos no son bajo ningún punto de vista "subjetivos", ya que existen independientemente de que sean parte de su experiencia del mundo. Están compuestos de materia, de células, y no existe algo así como una experiencia cualitativa de ser una célula, o en su defecto, tampoco hay algo así como una experiencia cualitativa de ser un libro. Estos objetos existen más allá de la experiencia personal que Ud. pueda tener.

¿Qué es lo que se intenta demostrar con esta distinción entre los elementos del parque? Se intenta probar que en el mundo existen nuestras experiencias subjetivas como algo concreto y real, y que simultáneamente, hay un mundo que existe más allá del hecho de que podamos experimentarlo. Esta dicotomía (X existe porque es experimentado por alguien v/s X existe más allá de las experiencias personales) lleva a una distinción entre lo mental y lo material. Según Searle (2004), esta diferenciación es la que constituye el problema mente-cuerpo, que según el autor, puede ser sintetizado en la forma de dos preguntas:

¿Cómo pueden existir las experiencias conscientes tales como el dolor en un mundo que se encuentra compuesto enteramente por partículas físicas?, y ¿cómo pueden algunas partículas físicas, ubicadas presumiblemente en el cerebro, causar las experiencias mentales? (p. 3, traducción mía.)

Puesto de otra manera, ¿Cómo puede lo mental, aquélla entidad física e incorpórea, afectar el mundo físico? E inmediatamente después, cabe preguntarse cómo lo mental posee un efecto sobre lo físico. Esta segunda interrogante es conocida en la literatura como el 'Problema de la causación Mental'.

Un tercer problema, al cual se le etiqueta como el 'Problema de la Intencionalidad'[6], puede ser hecho manifiesto en el ejercicio del lector en el parque, y referiría concretamente a los pensamientos que Ud. tiene sobre la economía nacional de los últimos meses. ¿Cómo pueden sus pensamientos "ubicados" en su cabeza, privados y personales, referir a objetos, entidades y estados distantes a su caja craneana? Es decir, ¿Cómo pueden los pensamientos ser *sobre* algo? Es esta pregunta la que corresponde al problema de la intencionalidad. En conclusión, la relación entre aquello que es *físico* y aquello que es *mental* posee un grado de complejidad que no refiere únicamente a las relaciones causales entre estos, sino también a otros aspectos de una dicotomía de esta naturaleza, lo que le entrega a la relación mente-cuerpo su estatus problemático.

1.2 Aproximaciones al Problema Mente-Cuerpo.

Históricamente, han existido dos escuelas de pensamiento con respecto al problema mente-cuerpo: El Dualismo y el Monismo. Si bien la mente comenzó a ser tópico de discusión ya desde los griegos (específicamente con Aristóteles), la Filosofía de la Mente en la era moderna comienza con René Descartes y su perspectiva dualista frente del problema. Sus argumentos fueron los más influyentes dentro de los filósofos modernos, particularmería desde la mitad del siglo XVII en adelante.

1.2.1 Dualismo.

La doctrina más famosa de Descartes es denominada 'Dualismo', y corresponde a la idea de que el mundo esta compuesto ontológicamente por dos tipos de sustancias[7]: Las Sustancias Mentales y las Sustancias Físicas. Cada una de estas sustancias tiene un rasgo esencial que la hace ser lo que es. La esencia de la mente sería la conciencia, mientras que la esencia del cuerpo sería su carácter "extendido", es decir, que se desenvuelve en un espacio físico tridimensional.

Al proponer que la esencia de la mente es la conciencia (o *res cogitans*), lo que Descartes quiere decir es que nosotros somos el tipo de seres que somos porque en todo momento nos encontramos en alguna forma de estado consciente, y es cuando dejamos de estar conscientes que dejamos de existir. Al plantear que la esencia del cuerpo es la extensión (*res extensa*), el autor refiere al hecho de que éste posee dimensiones espaciales, es decir, que tanto los objetos y entidades físicas del mundo, incluyendo al mismo mundo, se despliegan en el espacio.

Como seres humanos, en conclusión, somos entidades compuestas tanto por un cuerpo como por una mente, y esta entidad a la que referimos como 'Yo' correspondería

[6] Se hace referencia sucintamente a este problema en el Apéndice.

[7] Una sustancia se entiende como una entidad que tiene propiedades y que, a pesar de los cambios en sus propiedades, sigue siendo lo que es. Un perro es una sustancia, mientras que un huracán no lo es. Por lo tanto, decir que hay sustancias mentales significa que hay objetos que son no-materiales y que existen de manera independiente de lo físico.

a una mente que se encuentra adjunta a un cuerpo, y si bien están juntas, siguen siendo esencialmente distintas. Nuestra mente, de alguna manera, habita dentro de este cuerpo físico que poseemos, y esta mente no-física se encuentra en una interacción sistemática con el cuerpo que la contiene[8]. Por ejemplo, los estados físicos de los órganos sensoriales causan experiencias visuales, táctiles y olfativas en nuestra mente, así como los deseos y pensamientos de esta mente no-física hacen que el cuerpo se comporte de determinada manera. Esta influencia de la mente hacia el cuerpo se explica a través de una sustancia material sutil que él denominó 'Espíritus Animales'. Para Descartes, entonces, existen dos sustancias pero hay una predominancia de lo mental debido a que lo que podemos saber de manera directa y concreta es la vida mental, y accedemos a ésta a través de la introspección, lo que ratifica la naturaleza no-física de lo mental. Por otro lado, el conocimiento de lo físico, del cuerpo - de su existencia y características - funciona de manera distinta, ya que no puede ser conocido directamente, sino de manera indirecta a través de la inferencia de los contenidos de la propia mente. Es decir, no puedo percibir directamente el libro frente a mis ojos, sino más bien, percibo la representación del libro y no el libro mismo, e infiero que el libro existe dada la existencia de mi idea sobre él, ya que la idea de que tengo un libro en las manos debe ser causada por el libro.

Según Searle (2004), es importante enfatizar lo extremo de una doctrina como la de Descartes, y subrayar lo impactante de plantear que nuestro cuerpo físico no posee conciencia. No es la conjunción cuerpo-mente la que se encuentra en un estado consciente, sino sólo la mente. El mundo físico, ya sea una silla, una roca o el cuerpo humano, serían incapaces de poseer dichos estados, y es esta diferencia ontológica el principal rasgo distintivo en comparación con otras aproximaciones al problema, que si bien resulta intuitivo, es muy difícil de sostener científicamente.

1.2.2 Problemas del Dualismo Cartesiano y Un Nuevo Dualismo.

La doctrina cartesiana fue la más influyente durante los siguientes doscientos años. Pero ya desde autores clásicos como Hobbes y otros filósofos materialistas contemporáneos, existe un cuestionamiento a sus postulados y a la dualidad de sustancias. Lo anterior se produce debido a que lo planteado por Descartes adolecía de la capacidad de entregar una explicación adecuada a la relación entre cuerpo y mente. Si bien se establecía con claridad sus distintas características ontológicas e incorporaba la participación de Espíritus Animales como mediadores para salvaguardar la relación entre las substancias, el autor jamás refería a la manera en la cual los reinos de materia y mente se vinculan, interactúan o se modifican uno al otro, ya que, nuevamente, no se explica cómo aquello que es no-espacial puede interactuar con algo que sí lo es[9]. Y si bien el Dualismo otorgaba un cimiento filosófico que se adecuaba muy bien a las doctrinas religiosas de la época (una gran variedad de credos profesaban la propiedad de que el alma puede continuar viviendo una vez que el cuerpo

[8] Sin embargo, Descartes no es del todo claro con respecto a la manera en la cual estas dos sustancias ontológicamente distintas se vinculan, lo que de por sí es un aspecto problemático para el Dualismo Cartesiano. Más sobre esto en 1.2.2.

[9] Descartes no aclara la manera en la cual ambas sustancias interactúan. Sin embargo, sí refiere al lugar físico en el cual éstas interactuarían en el cuerpo. Dicho lugar sería en la Glándula Pineal, ubicada en la base del cerebro. El motivo para la elección de este lugar es que dicha glándula es la única zona del cerebro la cual no posee una división de dos partes o hemisferios, además de poseer una ubicación espacial en el centro ventral del cerebro. En los años en los cuales Descartes propusiera lo ulterior, no se sabía a ciencia cierta la función de dicha glándula; con el paso del tiempo se descubrió que ésta participa en la regulación de los ciclos de vigilia y sueño del organismo.

físico perece), la doctrina cartesiana comenzó a ser rechazada debido a su incapacidad de establecer relaciones causales claras entre lo físico y lo mental.

A modo de ejercicio y para ejemplificar lo extremo de una teoría como la dualista, podríamos realizar una reinterpretación de los postulados de Descartes a la luz de los descubrimientos de la Neurociencia en nuestros días; podríamos plantear la doctrina cartesiana utilizando un lenguaje actual o algo más científico, y quedaría de la siguiente manera: Nuestro cuerpo, los sistemas nerviosos y estructuras celulares no son ni pueden estar conscientes en ningún caso. El cuerpo y nuestro sistema nervioso central y autónomo serían contenedores, o bien escudos protectores que nuestra mente habita, y si bien se encuentran vivas tal como una planta se encuentra viva, el cuerpo en ningún caso se encuentra en un estado consciente, y nada de lo que la Física, la Neurobiología o la Inteligencia Artificial nos pueda decir arrojará luz sobre esta sustancia inmaterial. La mente se encuentra de alguna y misteriosa manera adjunta al cuerpo, y esta conexión cesa en el momento en que una persona muere, y una aproximación desde ciencias que estudian "la otra sustancia", no podría establecer de manera transparente cómo lo material explica la vida mental.[10] Nuevamente, a la luz de un vocabulario más actual, es posible entender lo drástica que suena una teoría como la cartesiana.

El problema de un planteamiento como éste es que resulta tremendamente problemático y difícil de implementar como hipótesis científica de trabajo. Se sabe que la actividad mental humana no puede desarrollarse ni producirse sin la existencia de ciertos tipos de procesos neurobiológicos del cerebro en particular, y del sistema nervioso en general. Quizás en algún futuro distante se pueda desarrollar conciencia en algún tipo de sustrato físico distinto al del sistema nervioso (eso sería materia de investigación empírica a futuro), pero en la actualidad no hay manera de saber si lo anterior es posible, ya que no contamos con una manera de acceder a la experiencia subjetiva de otros individuos. Por lo pronto, la neurobiología sugiere la existencia de una conexión evidente entre el cuerpo y la generación de estados mentales tales como la conciencia.

A modo de resumen, se puede decir que la fuerza del Dualismo cartesiano radica en tres tipos de argumentos. Primero, el 'Argumento Religioso', que encuentra fundamento en las doctrinas y creencias religiosas que proponen un tipo de vida después de la muerte y un tipo de yo no-corpóreo. Segundo, el 'Argumento de la Introspección', que refiere al hecho de que, como Descartes sugiriera, nuestra vida mental es a lo único que podemos acceder directamente y ergo, tiene preeminencia con respecto a lo extenso, que se conoce de manera indirecta. Y tercero, el 'Argumento de la Irreductibilidad', que consiste en negar el hecho de que las Ciencias Naturales puedan dar cuenta efectivamente de los fenómenos mentales que van más allá de lo físico, como la experiencia de degustar un vino (u otras experiencias sensoriales subjetivas en general), en la cual todo lo que un químico o un físico pueda saber en términos de estructura molecular, no le permitiría predecir ni anticipar dichas experiencias subjetivas. Estos tres argumentos, si bien pueden parecer atractivos, han sido criticados desde distintos puntos de vista que serán mencionados en la próxima sección.

Existe una versión posterior algo más moderada del Dualismo. Esta versión se etiqueta como 'Dualismo de Propiedades', en oposición al 'Dualismo de Sustancias' antes mencionado. El Dualismo de Propiedades difiere del Dualismo de Sustancia planteado por Descartes toda vez que no adscribe al hecho de que existan dos reinos separados, ni que mente y cuerpo sean sustancias distintas. Lo que plantea es que existen dos tipos diferentes

[10] Una interpretación cartesiana dejaría fuera cualquier tipo de aproximación en la cual el cuerpo participe activamente en la actividad cognitiva, i.e., la Cognición Corporalizada.

de *propiedades* (he ahí el origen de su nombre) en el mundo: Las Propiedades Físicas y las Propiedades Mentales[11]. Las primeras son aquellas propiedades que corresponden a cosas tales como la carga eléctrica, la masa de un cuerpo o la gravedad. Además, las propiedades físicas se caracterizan por ser mensurables, y caracterizan a un objeto o sistema como tal. Por lo tanto, cambios en las propiedades físicas de un objeto o sistema pueden ser usados para describir sus cambios. Las segundas corresponden a propiedades tales como la propiedad de sentir placer, la propiedad de tener un pensamiento sobre lo difícil que es escribir una tesis, o la propiedad de apreciar estéticamente una obra de arte; estas últimas serían las propiedades que son características de la inteligencia consciente. Un ejemplo concreto sería el siguiente: Si alguien masajea su cuello, Ud. probablemente se sentirá relajado y cómodo. Esa condición de sentirse relajado es una implementación de la propiedad de "Estar Relajado". El cuerpo humano, particularmente el cerebro, exhibiría ambos tipos de propiedades. Es decir, el Dualismo de Propiedades sostiene que no hay algo así como dos reinos ontológicamente distintos, sino que existen dos tipos de propiedades que son diferentes, las propiedades físicas y las propiedades mentales, y el cerebro posee la particularidad de poseer ambas.

El Dualismo de Propiedades posee otra característica que la diferencia del Dualismo de Sustancia: Si bien existen dos tipos de propiedades, existe un sólo tipo de sustancia, la sustancia física. Las propiedades mentales dependen tanto de las propiedades físicas y la sustancia física para existir. Sin embargo, algunas propiedades mentales no pueden ser "reducidas" explicativamente hablando a términos físicos, como por ejemplo las propiedades fenoménicas de la conciencia. Es esto último lo que diferencia crucialmente al Dualismo de Propiedades del Fisicismo.

Si bien el Dualismo de Propiedades salvaguarda la problemática de plantear la existencia de dos sustancias diferentes en el mundo, hereda del Dualismo de Sustancia el problema del vínculo mente-cuerpo antes mencionado, y no entrega respuestas alternativas en cuanto a la relación causal entre los eventos físicos y los eventos mentales. Más aún, surge un nuevo problema con respecto a cómo estas propiedades mentales funcionan causalmente en el mundo físico. El Dualismo de Propiedades no explica cómo funciona la causalidad de aquellas propiedades no físicas de mi cerebro (mis estados mentales conscientes) sobre los eventos que ocurren en el mundo físico. En otras palabras, la piedra de tope para este tipo de Dualismo puede ser resumida en la pregunta: ¿Cómo puede una propiedad mental jugar un rol causal sobre cualquier evento físico del mundo siendo que ésta no es física? Según lo demostrado por ciencias tales como la Física y la Química, el universo sería causalmente cerrado, i.e., la causación física sería suficiente para explicar cada uno de esos fenómenos del mundo. Por lo que sería posible concluir que los estados mentales conscientes no tienen ningún efecto en el universo, y podríamos establecer que cosas tales como la conciencia son algo así como un "epifenómeno" (Churchland, 1984), es decir, un fenómeno que si bien es producido por el funcionamiento del cerebro, es causalmente irrelevante. En este caso, la relación conciencia-cerebro sería similar al sonido que produce un computador en funcionamiento, que si bien se explica por el engranaje y conexiones de los componentes físicos, no tienen ninguna ingerencia en el resultado y procesamiento de las actividades de la máquina en cuestión. La conciencia y otros fenómenos mentales, en definitiva, serían un producto que emerge, y no tendría ningún efecto causal de retroalimentación sobre el sistema.

[11] Una propiedad se entiende como un atributo de un objeto o entidad; por ejemplo, de un objeto dulce se dice que tiene la propiedad de "Dulzura".

Por lo tanto, si bien el Dualismo de Propiedades no nos fuerza a creer que hay dos sustancias distintas en el universo, si lleva a postular que hay propiedades del cerebro que no son físicas y que son completamente distintas a todo el resto del ensamblaje bioquímico del cuerpo. Adicionalmente, no proporciona una respuesta para dar cuenta de estas propiedades en nuestra concepción del universo ni tampoco cómo éste funciona, lo que finalmente nos fuerza a seguir creyendo en la existencia de algo no-material e intangible al cual no podemos acceder más que a través de la introspección, sin ningún tipo de rigor científico.

Lo anterior conduce a optar por una de dos posibles alternativas: Aceptar el carácter epifenomenalista de nuestras sensaciones y experiencias del mundo, o simplemente considerar seriamente - utilizando la terminología de Descartes - la existencia de sólo una sustancia, en este caso, abrazar el postulado de que somos entidades compuestas por materia, o bien compuestos sólo por nuestra mente, lo que nos llevaría a una nueva forma de enfrentar el problema mente-cuerpo. Esta nueva aproximación se desarrolla a continuación: El Monismo.

1.2.3 Monismo.

Dados los problemas de una concepción del mundo en la cual hay dos tipos de sustancias o propiedades, la visión alternativa es que la citada dualidad no existe, y que sólo hay un tipo de sustancia, razón por la cual esta visión lleva por nombre *Monismo*. Ahora bien, dado que el Monismo plantea la existencia de una única sustancia, existirán dos tipos de Monismo: El Monismo 'Mentalista' o *Idealismo* y el Monismo 'Materialista' o *Materialismo*.

El Idealismo plantea que el universo es completamente mental o espiritual, es decir, que no existe nada más que *ideas* en el mundo, por lo que todo fenómeno mental y físico también es una idea. Es decir, se afirma que la naturaleza primordial de la realidad está basada en las ideas y la mente juega un rol como generador de éstas. En Filosofía de la Mente, el Idealismo funciona en oposición al Materialismo. El Materialismo propone que la naturaleza primordial de la realidad está basada en las sustancias físicas, y que no existe nada más que el universo material, por lo que toda explicación a un fenómeno es posible en términos materiales. De acuerdo a Blackmore (2005a), existe un problema con una visión "tan fuera de época" como la perspectiva idealista, dado que otorgar preponderancia a la mente y las ideas genera un conflicto con respecto a explicar cómo y por qué da la impresión de que efectivamente existe un mundo físico consistente al exterior de nuestra caja craneana. Al parecer, el mundo sigue existiendo sin mis ideas, y la gravedad, por más que ningún ser humano haya pensado en ésta o la haya entendido, sigue ahí. Es por eso que las perspectivas idealistas, si bien fueron relevantes desde un punto de vista histórico, carecen de importancia o ingerencia en el trabajo científico contemporáneo.

La familia de aproximaciones más influyente en la Filosofía de la Mente durante el siglo XX y lo que va del siglo XXI es alguna de las visiones del *Materialismo*. El Materialismo sostiene que la única realidad que existe es la realidad material. Nada existe o puede ser explicado por algo más que no sea la realidad física. Con respecto a los estados mentales, sostiene que si éstos efectivamente existen, entonces deben ser reducidos a estados físicos de algún tipo. Esta afirmación es la que se encuentra enraizada en la Ciencia Cognitiva desde sus inicios, así también en otras disciplinas que estudian la mente, proporcionando un ambiente fértil para la proliferación del su estudio, así como de otros fenómenos que conciernen nuestra vida mental.

Dado el problema de intentar dar cuenta de lo mental a través de lo físico, existen varias formas de Materialismo. Si bien todas éstas coinciden en la predominancia de lo físico como sustrato básico para explicar fenómenos mentales, algunas teorías enfatizan ciertos aspectos que otras visiones dejan de lado. Las aproximaciones a ser consideradas son: Conductismo, Materialismo Reduccionista (o Teoría de la Identidad), y Funcionalismo (específicamente, el Funcionalismo Computacionalista).

1.2.3.1 Conductismo.

La primera forma de Materialismo nació a principios del siglo XX, y comienza con el trabajo de Watson (1878 -1958) seguido por el de Skinner (1904 - 1990), alcanzando su nivel más alto de influencia durante las dos décadas posteriores a la Segunda Guerra Mundial. El planteamiento básico detrás del Conductismo nace, primeramente, como una reacción al Dualismo imperante hasta ese momento, y en segundo lugar, surge desde el Positivismo Lógico, el cual plantea que el significado de cualquier enunciado depende en última instancia de que puedan existir circunstancias observables *posibles* que lo confirmen. El argumento central del Conductismo establece que la mente corresponde únicamente al comportamiento del cuerpo, que no hay nada más allá de este comportamiento en términos objetivos, y aquello que constituye lo mental es precisamente este comportamiento, concretamente, la disposición a exhibir determinada conducta. Dicho de otra manera, se produce una reducción de lo mental sobre aquello que es posible observar, en este caso, la conducta. Ahora bien, es posible realizar una distinción entre dos tipos de Conductismo: El 'Conductismo Metodológico' y el 'Conductismo Lógico'.

El Conductismo Metodológico fue un movimiento que se desarrolló en Psicología, cuyo principal objetivo era otorgar a la disciplina un estatus científico respetable que le permitiera ser considerada dentro del resto de las Ciencias Naturales. Para lograr esto, propone que la Psicología debe estudiar aquello que es posible observar, y que, por tanto, sea objetivamente mensurable, razón por la cual se propone el estudio de la conducta. El objetivo de la Psicología sería entonces descubrir las leyes que pudieran proporcionar un correlato entre los estímulos de *input* que recibe el organismo con las respuestas de *output* conductuales.

El Conductismo Metodológico fue una perspectiva en Psicología de gran influencia, al punto que durante un tiempo logró posicionar una re-definición de la tarea de la Psicología, cambiando el enfoque de ésta desde la denominación de 'Ciencia de la Mente' a 'Ciencia de la Conducta'. El nombre de la aproximación - Conductismo Metodológico - se debe al hecho de que en ningún momento se buscaba hacer una afirmación o aclaración con respecto a la existencia o inexistencia de la mente ni determinar exactamente qué son los estados mentales, sino más bien proponer un método que le permitiera acceder a su estudio. En otras palabras, la mente se cataloga como científicamente irrelevante para el estudio de la conducta, ya que las afirmaciones científicas deben ser comprobadas de manera objetiva, y las únicas afirmaciones acerca de la mente humana son las afirmaciones que se pueden realizar con respecto a la conducta observable a través de un método objetivo. De hecho, el Conductismo Metodológico no necesitaba referir en ningún caso a la existencia de fenómenos mentales internos tales como la conciencia o los estados subjetivos o *qualia*, ya que el énfasis se encontraba en el posicionamiento de un método que iba más allá de la ontología de su objeto de estudio, y afirmaciones acerca de las emociones, sentimientos y creencias, eran sólo una manera abreviada de referir a patrones de conducta.

En resumen, este tipo de Conductismo sostiene que el único fenómeno psicológico que se puede observar en el ser humano es la conducta, por lo que el método apropiado para

la Psicología debe ser el estudio de la conducta humana sin prestar atención al estudio de entidades mentales o espíritus misteriosos.

También existió un movimiento conductista denominado Conductismo Lógico, pero éste emergió en la Filosofía, no en la Psicología. Los postulados del Conductismo Lógico eran más fuertes y radicales que los planteados por el Conductismo Metodológico; éste último planteaba que el Dualismo Cartesiano era científicamente irrelevante, mientras que el Conductismo Lógico postulaba que los planteamientos de Descartes estaban equivocados desde el punto de vista de la lógica y del análisis semántico, debido a que éstos serían pseudo-hipótesis y enunciados. Lo que esta afirmación sugiere es que un enunciado acerca de los estados mentales de una persona (ejemplo: “Marcelo espera que su equipo de fútbol gane un superclásico alguna vez”) pueden ser capturadas por un conjunto de oraciones acerca de las conductas (reales o posibles) a través de una definición operacional. Es decir, el enunciado debe ser traducible a un conjunto de oraciones hipotéticas sobre la conducta del individuo, en este caso, lo que Marcelo pudiera hacer o decir ante determinadas circunstancias.

En resumen, el Conductismo lógico sostiene que decir que “Luis cree que Colo-Colo es el mejor equipo del fútbol chileno” significa exactamente lo mismo que establecer un número indefinido de oraciones de causa-efecto tales como las que se sugieren a continuación: “Si Colo-Colo juega el domingo, Luis irá al estadio”, “si Colo-Colo hace un gol, Luis realizará una manifestación de júbilo”, y así sucesivamente, indicando que tener un estado mental en realidad es tener una serie de disposiciones a determinados tipos de comportamiento. Esta noción de disposición debe ser analizada usando oraciones hipotéticas del tipo “Si p, entonces q”, de manera de generar disposiciones conductuales con la forma “si tal condición/contexto/situación, entonces tal comportamiento se producirá”, por lo que aquellas expresiones que refieren a estados mentales pueden ser parafraseadas - sin ninguna perdida de contenido - en un enunciado lo suficientemente largo que de cuenta de cuáles conductas observables son esperables si la persona se encuentra en determinado contexto o circunstancia.

En conclusión, para el Conductismo el problema mente-cuerpo no es un problema en sí, porque según sus postulados, no tiene sentido preocuparse de una relación entre mente y cuerpo, ya que referir a la mente del individuo Einstein no es referir a “algo” que él “tiene”, sino que refiere a ciertos tipos de capacidades extraordinarias y disposiciones conductuales observables y posibles en torno al ambiente en que se encuentra situado el sujeto en cuestión.

Sin embargo, el Conductismo presentaba algunos problemas, que eventualmente lo dirigieron a un debilitamiento general. Primeramente, éste ignoraba deliberadamente, e incluso negaba, el aspecto interno o subjetivo de nuestros estados mentales, estableciendo que estos consistían en nada más que conducta y disposiciones conductuales. Sentir estrés, por ejemplo, es algo más que sólo andar con el ceño fruncido y tener problemas para dormir. Estar estresado posee una cualidad interna intrínsica a la cual sólo se puede acceder desde la introspección, e ignorar lo ulterior sólo denota la falta de capacidad del Conductismo de dar cuenta y explicar un fenómeno tan real como las conductas asociadas a este estado cualitativo interno. Simultáneamente, plantear la irrelevancia de los estados mentales va en contra de nuestra intuición de que hay una relación entre nuestros estados mentales internos y la conducta observable que se deriva de éstos, ya que existe una clara diferencia entre sentir dolor y las conductas asociadas a sentir un dolor. El segundo problema se produjo cuando los conductistas trataron de especificar detalladamente las disposiciones asociadas y constitutivas de una conducta: La lista de

condicionales necesaria para un análisis adecuado de “Marcelo espera que su equipo de fútbol gane un superclásico alguna vez” es demasiado larga, tanto que puede ser infinita, sin ninguna manera de especificar finitamente los elementos a ser incluidos, además del riesgo de incluir condiciones que apelen a estados internos del sujeto, es decir, la dificultad de evitar en sus definiciones palabras tales como “creer”, “querer” o “desear” para sólo utilizar circunstancias y conductas que sean públicamente observables, evitando circularidad en el proceso. Por ejemplo, la lista de condicionales necesaria para analizar “Rita desea comer chocolate” sería muy larga, tan larga que podría ser infinita, lo que complica y condiciona su análisis al ser poco específico y potencialmente interminable. Paralelamente, se corre el riesgo de definir una creencia en función de una conducta que puede estar asociada a otros deseos y creencias. Siguiendo el mismo ejemplo, suponiendo que Rita desea comer chocolate, el condicional (i) será cierto sólo si Rita aún no se *aburre* de salir a comprar a pie, y el condicional (ii) será cierto si *cree* que la heladería estará abierta en la tarde, y otras. Pero al ir haciendo una lista de condicionales que complemente una definición puramente conductual, se comenzaría a reincorporar una serie de elementos mentales (*aburre*, *cree*) en la definición, lo que significaría que ya no se define lo mental en términos públicamente observables exclusivamente. En tercer lugar, el Conductismo sufre una fuerte réplica por parte del lingüista Noam Chomsky, con respecto a la confusión del objeto de estudio de la Psicología. Chomsky afirma que es erróneo pensar que al estudiar la conducta se estudia Psicología, de la misma manera que sería equivocado sugerir que cuando se estudia la lectura de metros cuadrados se estudia Física. Estudiar la conducta, así como estudiar la manera cómo se miden los metros en la Física, correspondería al error de confundir la evidencia que se tiene sobre un objeto de estudio con el objeto de estudio mismo; aquél de la Psicología es la mente, y la conducta es la evidencia de la existencia de la mente. El Conductismo, según el autor, cometería este fatal error.

1.2.3.2 Teoría de la Identidad.

Cerca de la mitad del siglo XX, el Conductismo como proyecto metodológico en Psicología sufrió un fuerte revés debido a la serie de críticas antes presentadas. Es que al parecer, la aproximación conductista al problema mente-cuerpo parecía dejar algo fuera - la experiencia subjetiva. Además de carecer de una dimensión empírica científica que la respaldara, ésta posicionaba a la mente y la conducta dentro de un plano más bien lógico al apelar a componentes abstractos tales como listas y definiciones operacionales.

La cantidad de críticas y problemas que presentó el Conductismo hizo que fuera gradualmente sustituido por una nueva perspectiva entre aquellos filósofos que sostenían posiciones materialistas con respecto a la mente. Esta doctrina recibió el nombre de ‘Fisicismo’, también conocida con los nombres de ‘Teoría de la Identidad’ o ‘Materialismo Reductivo’ (en Churchland, 1984). El Fisicismo sostiene que Descartes comete un error al plantear la existencia de dos sustancias para explicar el problema mente-cuerpo, porque lo único que existe con respecto a la mente es el cerebro, y lo único que existe con respecto a los estados mentales son los estados del cerebro. Es por eso que la tesis del Fisicismo también es etiquetada como ‘Tesis de la Identidad’, ya que establece una relación de identidad e igualdad entre estados mentales y estados cerebrales. Los estados mentales son estados físicos del cerebro. Es decir, cada estado mental es numéricamente idéntico (es uno e idéntico al mismo tiempo) a un estado o proceso físico en el sistema nervioso, específicamente, en el cerebro.

La tesis central del Fisicismo contrasta radicalmente con el Conductismo en cuanto a la definición de un estado mental. Fundamentalmente, la Teoría de la Identidad proporciona

una perspectiva que no ignora la existencia de los estados mentales, sino que declara su existencia. Además, es una propuesta factual basada en hechos científicos concretos y que buscaba una reducción teórica al proponer los estados mentales como estados físicos, y no una propuesta lógica como la del Conductismo Lógico[12].

El Fisicismo sostiene y basa sus premisas en el hecho de que, tarde o temprano, la ciencia descubrirá que los estados mentales son idénticos a los estados cerebrales, realizando un caso de reducción inter-teórica (por ejemplo, véase en Churchland[13], 1984) como ha ocurrido históricamente en otras disciplinas: Así como se descubrió que el agua es idéntica al H2O, la ciencia logrará reducir los estados mentales a los estados cerebrales. La reducción inter-teórica recién mencionada se produce cuando una nueva teoría, lo suficientemente poderosa en términos explicativos y predictivos, termina por incluir un grupo de proposiciones y principios que reflejan a la teoría antigua casi de manera perfecta. Históricamente se pueden mencionar casos en los cuales nociones pertenecientes a una teoría "antigua" (por ejemplo, "calor" corresponde a "X está caliente" o "X está frío") se incluyen en la nueva teoría ("energía cinética molecular total", "el objeto posee un promedio alto de energía cinética molecular" y "el objeto posee un promedio bajo de energía cinética molecular", respectivamente), lo que permite generar identidades inter-teóricas. Por tanto, la Teoría de la Identidad sostiene que, eventualmente, la ciencia (particularmente la Neurociencia) generará una teoría la cual dé cuenta del hecho de que un estado mental corresponde a un estado cerebral.

Las razones para sostener una teoría materialista-reductiva pueden ser resumidas en cuatro ideas básicas que tienen como premisa central que la conducta humana y sus causas encuentran su campo de estudio en las Neurociencias. El primer argumento reside en el origen puramente físico y la constitución física del ser humano. El desarrollo físico-motor del hombre puede ser explicado en términos de moléculas, duplicación celular, y cadenas de ADN que portan la información para desarrollar ciertos procesos y conductas de un sistema, conductas que a su vez se encuentran interaccionando constantemente con el ambiente. El segundo argumento se encuentra conectado con el primero, y hace alusión al hecho de que otros animales en la naturaleza también encuentran su origen y desarrollo en una dimensión física, específicamente, en el funcionamiento adecuado del sistema nervioso, y su conducta también encuentra explicación en los procesos internos del sistema y su interacción con el ambiente; lo anterior tiene origen en la teoría evolutiva, la cual sería responsable de la selección de los sistemas antes mencionados. Tercero, los teóricos adhieren a la idea de una "dependencia neuronal" de los fenómenos mentales, que resulta fundamental para una premisa materialista-reduccionista, en la cual se descarta la existencia de dos tipos de sustancia, o propiedades o procesos ya que sólo lo físico es relevante. El cuarto argumento hace alusión al éxito de la Neurociencia en la investigación y descubrimiento de la estructura de los diversos sistemas nerviosos tanto de criaturas simples así como la del propio ser humano. Es decir, dado el rápido desarrollo de la disciplina, es esperable que nuevos y clarificadores descubrimientos se puedan producir en los años por venir. Los teóricos de la identidad aseguran que es razonable creer que lo que causa la conducta animal y humana es esencialmente físico, pero no sólo eso, también aseguran que la Neurociencia

[12] Como se mencionara en la sección anterior, el modelo propuesto por el Conductismo Lógico era de identidades definicionales, en la cual sensaciones tales como "placer gustativo" no serían más que disposiciones a una determinada conducta, definiendo por tanto un estado mental de la misma manera en la cual se define un cuadrado, por ejemplo, como una figura bidimensional de 4 lados iguales.

[13] La cita proviene de un libro escrito por Churchland en el cual expone sobre la teoría reduccionista, pero este autor no adscribe a una teoría reduccionista de la mente.

desarrollará una taxonomía tal de los estados cerebrales, que se podrá desarrollar una relación *uno a uno* con los estados mentales conscientes.

Lo propuesto por la Teoría de la Identidad es un avance con respecto al Conductismo, ya que efectivamente la teoría se encarga de entregar una propuesta con respecto a qué son los estados mentales, en vez de ignorarlos como su predecesora. Sin embargo, una perspectiva como ésta presenta una serie de objeciones y problemas, las cuales de acuerdo a Searle (2004) pueden ser clasificadas de la siguiente manera: Las 'Objeciones Técnicas' y las 'Objeciones de Sentido Común'.

Dentro del grupo de las objeciones técnicas se puede señalar principalmente que la Tesis de la Identidad viola el principio lógico denominado "Ley de Leibniz" (Searle, 2004, p. 39). La ley establece, en muy pocas palabras, que si dos cosas son idénticas, entonces deben compartir las mismas propiedades. Por lo que si se puede demostrar que los estados mentales poseen propiedades que no pueden ser atribuibles a los estados cerebrales, y/o si los estados cerebrales tuvieran propiedades que no pudieran ser atribuidos a los estados mentales, se podría refutar la Tesis de la Identidad. Y la Teoría de la Identidad es susceptible a tal observación, comúnmente citando casos de localización espacial de partes de nuestro cuerpo. Por ejemplo, supongamos que el estado cerebral que corresponde a mi pensamiento de que hace calor se encuentra *localizado* cerca del área 48 en mi cerebro. Sin embargo, los críticos de la Teoría de la Identidad afirman que no tendría sentido asegurar algo así como que mi pensamiento de que hace calor está "localizado" en el algún punto particular de mi cerebro. Lo anterior incluso carece de sentido en el caso de los estados mentales conscientes que poseen una *ubicación espacial* en el cuerpo, tales como la comezón en el pulgar derecho de mi mano; en este caso, se afirma que efectivamente la comezón puede estar en mi pulgar derecho, pero el estado cerebral que corresponde a la citada comezón no está en mi pulgar, sino que en otro lugar, en mi cerebro, lo que lleva a concluir que las propiedades del estado cerebral no son iguales a las propiedades del estado mental, por lo que el Fisicismo sería falso[14].

Una segunda objeción técnica es la citada por Churchland (1984), y puede ser etiquetada como la crítica a las 'Propiedades Semánticas del Cerebro'. El argumento consiste en lo siguiente: Nuestras creencias, deseos, etc. poseen un significado, un significado proposicional específico, y pueden ser verdaderos o falsos, sosteniendo relaciones semánticas de consistencia, implicaturay otras entre éstas.Ahora bien, si los estados mentales fueran sólo estados cerebrales, entonces todas estas relaciones entre creencias, deseos y pensamientos debieran ser ciertas también para los estados cerebrales. Sin embargo, carece de sentido decir que la actividad neuronal en mi corteza temporo-parietal es falsa, o que implica lógicamente a otra área de la corteza, o que tenga el significado que X. En otras palabras, aún parece lejano o al menos difícil otorgar propiedades semánticas a los estados cerebrales. Más complejo aún sería asumir una posición en la cual se pueda establecer una relación de identidad entre estado cerebral X y la experiencia subjetiva del mundo, por ejemplo, una relación de identidad entre el estado cerebral X y la sensación subjetiva de masticar algo crujiente.

El segundo tipo de objeciones de acuerdo con la clasificación de Searle consistiría en aquéllas que provienen desde el sentido común. La más importante objeción de sentido común consistiría en demostrar que si efectivamente la identidad estado-mental / estado-

[14] Ahora bien, desde el Fisicismo existe una manera de responder a esta objeción, y la respuesta consistiría básicamente en que la Teoría de la Identidad no se encuentra interesada en un objeto putativo, en este caso, una comezón, sino más bien en la experiencia completa de sentir comezón, y dicha experiencia se extiende desde la misma terminación nerviosa en mi pulgar hasta el cerebro. De acuerdo a Searle (2004), la respuesta es lo suficientemente contundente como para dar cuenta de la objeción.

cerebral que se propone es empírica (de la misma manera como se plantea que el agua es H2O), entonces se está sugiriendo que existirían dos tipos de propiedades que sirven de puente entre los dos aspectos de la identidad. Es decir, de la misma manera en la cual la Teoría de la Identidad sostiene que el agua es idéntica con H2O y se debe identificar al objeto en cuestión tanto en términos de sus propiedades de "agua" como en términos de sus propiedades de "H2O", ésta debiera explicar que la sensación gustativa dulce es idéntica con el estado cerebral X tanto en términos de las propiedades de sensación gustativa dulce así como en términos de las propiedades de los estados cerebrales X correspondientes a aquella sensación. Sin embargo, un planteamiento en el cual se presentan dos tipos de propiedades distintas e independientes pareciera sugerir que existen propiedades físicas y propiedades mentales, lo que se asemeja peligrosamente a un tipo de Dualismo de Propiedades; es decir, si los estados mentales son estados cerebrales, entonces hay dos tipos de estados cerebrales, aquellos que son estados mentales y aquellos que no son estados mentales, en los cuales los primeros poseen propiedades mentales y los segundos tienen propiedades físicas.

La objeción de sentido común recién planteada representa un obstáculo concreto en la argumentación de los que apoyaban la tesis de la Teoría de la Identidad. El obstáculo radica en que si bien ésta buscaba establecer una reducción que implicaba la idea de que los estados mentales son única y exclusivamente estados cerebrales, la tesis deja entrever la existencia de propiedades mentales que son irreductibles. Al parecer, es imposible dar cuenta de la rica experiencia del individuo consciente con respecto a sus penas, alegrías y sensaciones corporales a través de la reducción de un fenómeno al lugar físico dónde éste se produce. Es decir, la reducción se encarga de estandarizar ciertas características a través de la adjudicación de propiedades de la experiencia subjetiva a las propiedades del cerebro, sin embargo, la reducción no resulta ser exitosa al momento de intentar dar cuenta del estado cualitativo y subjetivo de esos determinados fenómenos mentales ignorando (y hasta cierto punto, eliminando) la existencia de propiedades mentales no-físicas.

Un segundo problema de la Teoría de la Identidad nace desde lo que podría etiquetarse como "Chauvinismo Neuronal". La crítica sostiene que si toda experiencia subjetiva es idéntica a un tipo determinado de estimulación neuronal, y si cada creencia es idéntica a un tipo especifico de estado neuronal, entonces pareciera ser que una criatura que no tenga una constitución neurológica como la nuestra - o que al menos sea lo suficientemente parecida - no podría experimentar creencias, deseos, comezón, placer, etc. Lo que el argumento pareciera posicionar es una crítica a algo así como la "supremacía neuronal" en el argumento del Fisicismo, al entregar una relación de identidad inseparable entre la experiencia subjetiva de un ser vivo con su sustrato neuro-fisiológico.

La anterior objeción produjo un impacto real en la Teoría de la Identidad, generando una distinción antes/después en ésta que se hizo efectiva en un cambio de nombre de 'Teoría de la Identidad *type-type*' a 'Teoría de la Identidad *token-token*'[15]. La etapa previa de la teoría (*type-type*) establecía que cada *type* de estado mental es idéntico con un *type* de estado físico. Pero si se hace un análisis más cuidadoso, es posible ver que tal afirmación no es correcta, ya que la relación de identidad que se produce en la vida cotidiana es, de hecho, una relación entre realizaciones reales, situadas en el mundo, concretas (i.e., *tokens*), y no entre conceptos abstractos carentes de temporalidad y abstraídos del espacio físico y el contexto. Esta aparentemente ligera observación en definitiva hizo que algunos de los teóricos de la identidad adoptaran una posición *token-token* que propone

[15] La distinción *type-token* puede ser entendida considerando a un *token* como las realizaciones o ejemplificaciones particulares concretas de un concepto abstracto tipo. Por ejemplo, los estadios Santa Laura y Monumental serían *tokens* del *type* Estadio de fútbol.

lo siguiente: No todo *token* de dolor debe ejemplificar o realizar el mismo *type* de estado cerebral; pueden haber *tokens* de distintos *types* de estados cerebrales, aunque todos sean *tokens* del mismo *type* de estado mental, en este caso, el *type* de dolor. En otras palabras, para cada *token* de un estado mental X existe un *token* de un estado físico Z con el cual es idéntico. Por ejemplo, supongamos que el Presidente de Chile cree que la economía del país está avanzando, y supongamos que el líder de la oposición política también cree que la economía del país está avanzando. Que ambos sujetos tengan la misma creencia hace innecesario suponer que para que ellos exhiban esa creencia deban compartir todos y cada uno de los rasgos neurobiológicos del estado cerebral en cuestión. Dicha creencia puede generarse en la interacción de ciertas áreas del cerebro del Presidente que no sean exactamente las mismas áreas en el líder de la oposición, pero eso no impide que tengan la misma creencia.

Si bien el movimiento desde una Teoría de la Identidad *type-type* a una teoría *token-token* daba cuenta de ciertos problemas en cuanto a la realización de los fenómenos mentales, se produce un problema definitivo que finalmente llevaría a cuestionar seriamente la plausibilidad de una identidad reductiva de la experiencia subjetiva a los estados cerebrales. La pregunta es: ¿Qué hace a todos los *tokens* ser un *token* de un determinado *type* de estado mental*?* Es decir, ¿Qué hace a un *token* X ser la instanciación o ejemplificación de un *type* mental X? La Teoría de la Identidad se ve incapacitada de dar cuenta de este problema. La respuesta a esa pregunta dio origen a una nueva manera de aproximarse al problema: Un *token* es el *token* de determinado *type* mental debido a la *función* que cumple en el comportamiento de una criatura. A dicha aproximación se le denominó 'Funcionalismo'.

1.2.3.3 Funcionalismo.

Dado los problemas anteriormente presentados con respecto a la Teoría de la Identidad, surge una nueva perspectiva con respecto a los estados mentales, la cual introduce una manera distinta de comprender y definir un estado mental sin presentar los problemas que otras teorías exhibían. La nueva aproximación recibe el nombre de 'Funcionalismo', la cual se yergue sobre la idea de que un *token* de estado cerebral constituye un estado mental propiamente tal si es que éste posee una determinada *función* en la conducta del organismo al cual pertenece.

De acuerdo con el Funcionalismo, lo que define de manera esencial a un estado mental es el conjunto de relaciones causales que éstos tienen con (1) las interacciones del ambiente con el cuerpo, (2) otros estados mentales y (3) la conducta del cuerpo. Es decir, es posible identificar a un dolor como tal si es que (1) es producido por algún tipo de lesión o herida en el cuerpo, (2) produce una sensación de incomodidad y molestia, además de inducir una actitud que permita aliviarlo y paralelamente, (3) estremecimientos, potenciales mareos y especial tendencia a proteger el área afectada. El Funcionalismo establecería que si existe un estado mental que cumpla exactamente las funciones recién mencionadas, entonces éste sería un dolor.

Bajo esta nueva premisa, cualquier tipo de estado mental puede ser definido en términos de sus roles causales específicos y únicos que actúan como mediadores entre los *inputs* sensoriales y los *outputs* conductuales. Por ejemplo, considere la siguiente situación: Según el Funcionalismo, decir que Luis cree que el vino produce malestares estomacales, es decir que existe un proceso o estado que ocurre dentro del sujeto Luis causado por ciertos estímulos externos (por ejemplo, cada vez que el sujeto Luis ha ingerido el líquido en cuestión ha sufrido dolores y calambres en la zona abdominal), y que esto

último sumado a otros factores tales como el instinto de rehuir a los malestares que los seres humanos poseemos, producirán como resultado cierto tipo de conducta, en este caso, evitar beber vino en la próxima comida que tenga con sus padres. Por lo tanto, la creencia de que el vino produce malestares estomacales, sumado al instinto humano de evitar estados poco placenteros, producen causalmente la conducta de no beber vino en las comidas con sus padres. En definitiva, los estados mentales son entonces definidos como estados que tienen determinadas funciones, siendo estas funciones entendidas como relaciones causales entre estímulos externos y conducta observable, lo que significa que los estados mentales no son definidos por características intrínsecas internas, sino que por estas relaciones causales las cuales constituyen, en definitiva, su función.

Una perspectiva que establece relaciones causales entre *inputs* sensoriales y *outputs* conductuales trae a la memoria una cierta proximidad con los planteamientos del Conductismo descritos con anterioridad (ver sección 1.2.3.1). Y claro, en ambos se hace énfasis en el rol fundamental del ambiente como generador de la conducta. Sin embargo, existen algunas diferencias que adquieren relevancia al momento de, efectivamente, realizar una comparación, y que permite distinguirlos claramente. La diferencia radica en el hecho de que los conductistas buscaban definir cada estado mental utilizando básicamente términos de *inputs* y *outputs* externos. El Funcionalismo dista de plantear cosa semejante; por el contrario, la caracterización de un estado mental requiere imperiosamente la existencia de otros estados mentales con los cuales se encuentre conectado causalmente, dejando de lado cualquier intento de reducir lo externamente observable a los procesos mentales. En otras palabras, si bien existe una similitud con respecto al rol crucial que juega el ambiente en la generación de conductas tanto en el Conductismo como en el Funcionalismo, este último establece que en ningún caso el contexto sea lo suficientemente explicativo para dar cuenta de la conducta. El Funcionalismo entrega un rol preponderante a la mente en la generación de esta última, a través de las relaciones causales de los estados mentales tanto con el ambiente como con otros estados mentales.

Es interesante prestar atención al hecho de que el Funcionalismo posee una característica que ninguna de las teorías anteriormente expuestas exhibe, y que quizás le entregó una gran popularidad entre los teóricos materialistas. Su mayor logro es la presentación de lo mental es términos funcionalmente concretos, similares a otros dominios del conocimiento humano, en este caso, logrando que la mente pueda considerarse como una máquina. Por ejemplo, cuando referimos a una lámpara, el Funcionalismo plantea que, independientemente de los elementos físicos que la constituyan o de los posibles mecanismos y materiales que utilicemos para diseñar una (por ejemplo, utilizar una lámpara en función de reacciones químicas, utilización de ampolletas o una bolsa llena de luciérnagas, y sea esta grande, pequeña o alargada), lo que hace a una lámpara ser tal artefacto es que genera luz; lo que define a la lámpara es su función, y no su estructura física ni su esencia cartesiana. Los estados mentales son sólo una entidad que, dadas ciertas relaciones con el ambiente y otros estados mentales, producen conducta. Hasta ahí, se podría decir que el Funcionalismo es exitoso en sus planteamientos, toda vez que proporciona una respuesta satisfactoria con respecto a qué es un estado mental, entregando una explicación concreta y reconocible, y qué incluso, parece intuitivamente acertada.

Según Churchland (1984), el Funcionalismo es la teoría que goza de mayor aceptación entre los filósofos de la mente, cientistas cognitivos e investigadores en Inteligencia Artificial. Dicha aceptación proviene del hecho de que una perspectiva como ésta permite separar función de implementación, lo que hace posible caracterizar la mente en

términos funcionales con independencia de la constitución física del cerebro. Es decir, una concepción funcionalista de la mente permite abstraer a la mente de una dependencia al sustrato físico sobre el que opera (Block, 1995a), y es ésta una de las principales razones para la aceptación general que Churchland comenta. Simultáneamente, el Funcionalismo permitió a la Psicología adquirir autonomía metodológica de disciplinas tales como la Filosofía o la Neurociencia. Lo anterior permite sustraer a los estados mentales de terrenos científicos los cuales o bien los ignoran, o los reducen al sustrato físico en el cual se implementan. Un tercer motivo sería que el Funcionalismo permite arrojar luz sobre la actividad cognitiva de otros seres vivos tales como la de los animales, y al mismo tiempo expandir el debate sobre la posibilidad de existencia de actividad inteligente en programas computacionales y eventualmente, en criaturas artificiales, lo que contribuye a caracterizar la inteligencia desde un punto de vista más universal.

En el siguiente capítulo exploraremos como el funcionalismo permitió el surgimiento de una nueva aproximación al problema mente-cuerpo, aquélla que posicionaba la metáfora de que la mente es un computador.

2. Una Nueva Respuesta al Problema Mente-Cuerpo: Ciencia Cognitiva.

2.1 Funcionalismo y Ciencia Cognitiva.

A mediados del siglo XX comienza a gestarse uno de los hitos históricos más importantes en el estudio de la mente. Este hito se origina en la convergencia del trabajo de múltiples disciplinas involucradas en su estudio, entre las cuales se puede contar el trabajo de la Filosofía de la Mente, la Psicología Cognitiva, la Lingüística y la Inteligencia Artificial (I.A.). El trabajo interdisciplinario entre éstas ha llevado a una nueva aproximación al problema-mente cuerpo. Dicha aproximación posiciona una idea novedosa inspirada en una teoría funcionalista de la mente, la cual permite introducir una manera alternativa de comprender su funcionamiento, y de paso, otorgar una visión diferente al problema que por siglos mantuvo ocupados a filósofos y psicólogos. Esta idea novedosa consistiría en que la mente es un computador.

Dado el advenimiento de la computación durante el periodo de la post-guerra a mediados del siglo pasado, y dada la importancia del trabajo de Alan Turing (1912 - 1954) y sus conceptos de 'Computación' y 'Algoritmo', comienza a considerarse el argumento de que el cerebro es un computador digital. En este argumento, la analogía corresponde a que la mente es el *software* computacional o programa (quizás un conjunto de programas) que opera en el cerebro, en la que los estados mentales comienzan a ser entendidos como estados computacionales del cerebro. En otras palabras, a través de la Ciencia Cognitiva[16], se introduce una nueva aproximación para la comprensión de los fenómenos mentales, la cual establece que la mente es al cerebro lo que el *software* es al *hardware* computacional. Esto es lo que se podría denominar la visión 'Computacionalista de la Mente'.

Antes de continuar, es pertinente presentar la siguiente pregunta: ¿Qué motiva el surgimiento de esta visión computacionalista de la mente? Para poder responder esto, es necesario antes reunir ciertos datos e información que expliquen el surgimiento de la Ciencia Cognitiva, lo que nos entregará el contexto necesario para comprender y aproximarnos críticamente a una teoría computacionalista de la mente.

2.2 Un Poco de Historia: Ciencia Cognitiva.

La Ciencia Cognitiva surge históricamente como una reacción a sus dos antecedentes disciplinarios, la Psicología Introspectiva y el Conductismo. La Psicología Introspectiva

[16] Evidencia de la conexión entre la Ciencia Cognitiva y el paradigma que el cerebro es un computador digital se puede encontrar en una entrevista dada por Turing (en Jones, 2004) en 1951, y en ésta podemos ver como se forjó la idea de que un computador digital tiene estados mentales y el cerebro puede comprenderse como un computador programado. Turing señala que 'Si se acepta que los cerebros reales, tales como los encontrados en los animales y particularmente en los hombres, son un tipo de máquina, se sigue que un computador digital bien programado, se comportará como un cerebro" (p. 9, traducción mía).

surge durante el siglo XIX, y su autoría se le otorga al psicólogo alemán Wilhelm Wundt (1832 - 1920). Esencialmente, la Psicología Introspectiva sostiene que el acceso a la conciencia es posible utilizando el auto-análisis como método de introspección. Éste consistía en entrenar sujetos para que fueran capaces de analizar sus estados conscientes y dar cuenta de ellos a través de un lenguaje determinado, otorgando una preponderancia a la vida mental en detrimento de la observación externa de la conducta (Bechtel, 1998). En definitiva, el gran mérito de Wundt radica en el hecho de entregar un cariz científico a la Psicología a través de sus postulados.

La corriente de pensamiento que le sigue es el Conductismo, y se caracteriza principalmente por realizar una aseveración proporcionalmente opuesta a la de la Psicología Introspectiva: El estudio de la mente se encuentra remitido a la observación sistemática de aquello que puede ser mensurable desde la tercera persona, i.e., la conducta. El estudio de lo mental adquiere el perfil de una teoría de la conducta, y si bien lo mental existe, es explicativamente irrelevante debido a la imposibilidad de acceder objetivamente a ésta. Es decir, el Conductismo tiene como premisa realizar un estudio de aquello que pueda ser medido y, por lo tanto, sistematizado, en desmedro de una aproximación que dependa de la subjetividad del investigador para obtener resultados. He ahí la dificultad de estudiar un fenómeno subjetivo como la conciencia, ya que su existencia se produce en la medida en que es experimentada por un observador, lo que la hace imposible de generalizar y definir de manera objetiva.

Por lo tanto, el contexto en el cual comienza a desarrollarse la Ciencia Cognitiva es el contexto que imperaba desde los planteamientos del Conductismo: Los estados mentales que usualmente asociamos a nuestra vida mental como pensamientos, sentimientos, emociones, tienen componentes que no son observables directamente y cuya existencia no queda reflejada de manera fidedigna por lo conductual. Y ese fue, precisamente, el desafío para los cientistas cognitivos: Generar un marco teórico capaz de lidiar con la inaccesibilidad a los fenómenos mentales desde la tercera persona. La manera de lograrlo fue a través del modelamiento de la mente como un computador. A través de la metáfora del computador es posible obtener un modelamiento que entregue la posibilidad de ser probado e implementado, además de poder ser manipulado para obtener datos concretos en cuanto a su funcionamiento. Paralelamente, el modelamiento de una mente puede ser llevado a cabo al diseñar programas y hacerlos funcionar en el computador, para así poder describir su funcionamiento e implementación de manera objetiva.

2.2.1 Un Alto en el Camino: Turing y Algunas Nociones Básicas de la I.A.

Para poder explicar las ventajas de una teoría computacionalista de la mente, es fundamental manejar ciertas nociones básicas que permitan entender por qué algunos filósofos plantean que esta aproximación es la mejor respuesta al problema mente-cuerpo. Dichas nociones se encuentran conectadas entre sí, y son las siguientes: Algoritmo, Máquina de Turing, Tesis de Church y el Test de Turing.

Primeramente, la noción de ‘Algoritmo’ resulta fundamental. Un algoritmo se entiende como un procedimiento efectivo, es decir, una serie de pasos finitos cuya ejecución lleva a la resolución de un problema. Estos pasos son operaciones simples, precisas y finitas, que se realizan mecánicamente y no requieren de la participación de la conciencia. Un algoritmo, dada su naturaleza, puede ser llevado a cabo tanto por un computador complejo, así como por una persona equipada con lápiz y papel, ya que para que un algoritmo funcione, no

se necesita nada adicional a las operaciones que éste trae incorporado. Existe un grupo de *tokens* iniciales que funcionan como *input*, existe el estado final o resultado al que se llega o *output*, y las operaciones que se ejecutan sobre los *inputs* corresponden a estados de transición que van modificando los *tokens* iniciales. A estas operaciones también se les denomina 'Computaciones', y se encuentran en un estado u otro. El ejercicio clásico de Euclides para calcular el máximo común divisor (MCD) de dos números es un buen ejemplo de un algoritmo:

Paso 1. Dados dos números positivos m y n, sea m mayor que n. Paso 2. Dividir m por n. Guardar el remanente como r. Paso 3. Si r=0, entonces parar; el MCD es n. Paso 4. Si no, ponga n en el lugar donde iba m, y ponga r en la posición de n. Entonces, vaya al paso 2.

La definición de una función computacional se conecta directamente con el segundo concepto a presentar, aquel denominado 'Máquina de Turing'.

Una Máquina de Turing es un modelo matemático de una máquina que opera de manera mecánica, que funciona leyendo símbolos para luego seguir escribiendo otros a través del uso de un cabezal. Todo el proceso se encuentra guiado por una tabla de instrucciones o programa. La operación de la máquina se encuentra totalmente determinada por esta lista finita de instrucciones simples (ejemplo: "En el estado 97, si lee el símbolo 8, escriba 1; si lee el símbolo 9, muévase hacia la derecha, y vaya al estado 7"). La máquina planteada por Turing es una idealización matemática sobre qué es computar, y jamás fue pensada como una tecnología computacional real, sino más bien como un experimento mental que representa una máquina que computa (en la versión original, los pasos eran seguidos, de hecho, por una persona real que el autor etiquetaba como "computador"). Concretamente, y de acuerdo al autor, una Máquina de Turing opera de la siguiente manera:

[Una máquina de Turing tiene] una capacidad de memoria infinita adquirida en la forma de una cinta infinita dividida en cuadros, en los cuales se pueden imprimir símbolos. En determinado momento, habrá un símbolo en la máquina, y éste se denomina símbolo escaneado. La máquina puede alterar el símbolo escaneado, y su conducta se encuentra determinada, en parte, por ese símbolo, pero los símbolos que se encuentren en el resto de la cinta no afectan la conducta de la máquina. Sin embargo, la cinta puede moverse para adelante y para atrás a través de la máquina, siendo ésta una de sus operaciones básicas. Cualquier símbolo en la cinta puede, eventualmente, convertirse en el símbolo de turno que la máquina escanea. (Turing, 1948, p. 61, traducción mía.)

El dispositivo se caracteriza por operar con símbolos en binario y por esa razón utiliza típicamente 1s y 0s, aunque puede utilizar otros símbolos. Lo que caracteriza más profundamente a esta máquina es su particular simplicidad, ya que las reglas siempre tienen la forma de "bajo condición X, realice acción A: C-->D", que si bien es simple, sintetiza y delimita el concepto de computación mecánica, en la cual no se necesita ningún tipo de entendimiento adicional y/o especial para llevar a cabo la tarea. Finalmente, una Máquina de Turing se caracteriza por tener estados internos, y la conducta de ésta queda determinada mediante pasos necesarios. En definitiva, es posible decir que una Máquina de Turing es, de hecho, equivalente a la tabla de máquina o lista de instrucciones que determinan todos los estados y procesos en los que ésta puede encontrarse.

El tercer concepto, la 'Tesis de Church', se relaciona con los dos conceptos previamente expuestos, y de manera simple y resumida, establece que cada computación efectiva puede ser llevada a cabo por una Máquina de Turing. Lo anterior también se conoce como la

'Tesis de Church-Turing', ya que Turing llega a una conclusión similar, sólo que de manera independiente y posterior a Alonzo Church (1903 - 1995). Esta tesis se vincula con la noción de un procedimiento efectivo o mecánico M que conduce a un resultado deseado. Éste posee ciertas características (Copeland, 1993), entre las cuales destacan poseer un grupo de instrucciones finitas, la posibilidad de lograr el resultado deseado si es que se ejecuta sin errores y que puede ser llevado a cabo por una máquina o persona sin la necesidad de utilizar conciencia o algo ajeno al programa en sí. La tesis, por lo tanto, afirma que cada vez que haya un método efectivo que utilice esos procedimientos mecánicos para obtener los resultados deseados, éstos pueden ser computados por una Máquina de Turing, ya que ésta es en sí misma una especificación de lo que es un método efectivo: La utilización de estados iniciales e intermedios que junto a la manipulación de reglas y símbolos, permite llegar a un estado final sin necesidad de nada más que el programa. En resumen, sería posible decir que cualquier computación o cálculo puede ser traducida a su computación equivalente en una Máquina de Turing[17]; es decir, una Máquina de Turing agota completamente la idea de computabilidad, ya que todo algoritmo puede ejecutarse por una Máquina de Turing, y dado que la conducta de ésta siempre es algorítmica, permite establecer una conexión entre una definición formal con una intuitiva, a saber, se identifica una clase de funciones definidas formalmente (las funciones recursivas) con las funciones que son computables, i.e., aquellas para las cuales se puede escribir un algoritmo.

Es de suma importancia destacar la importancia de la tesis y es por eso que nos detendremos en esto por un momento. La idea de una máquina que ocupa elementos muy básicos y simples - ceros y unos entendidos como símbolos y las reglas de transformación - pueda ser capaz de llevar a cabo cualquier tarea que requiera un algoritmo, significa que puede haber una Máquina de Turing para realizar distintas tareas que requieran de distintos algoritmos, tales como una máquina que mida la temperatura del aire en una habitación, un cajero automático, una máquina que permita entregar una bebida cada vez que se le inserta una moneda, etc. Pero además de la posibilidad de que existan computadores para tareas específicas basadas en algoritmos (es decir, una Máquina de Turing), es plausible considerar la existencia de una máquina que sea capaz de implementar los programas de otras Máquinas de Turing, es decir, una Máquina de Turing para propósitos generales a la que se le pueda "cargar" los programas de otras Máquinas de Turing. Esta idea proviene de lo que se conoce como 'Máquina Universal de Turing' la cual puede simular el funcionamiento de otras Máquinas de Turing (Turing, 1948).

El planteamiento básico de la Tesis de Church funciona como un ingrediente crucial para la investigación de la mente y el uso de máquinas que computen para simular aspectos de la mente humana, ya que es posible considerar a esta última como una Máquina Universal de Turing. Lo anterior resulta ser trascendente tanto para la Inteligencia Artificial como la Psicología Cognitiva, dado que entrega una noción teórica básica y común que permitiría una aproximación conjunta por parte de la Ciencia Cognitiva al problema mente-cuerpo, y a su vez, permite estudiar el funcionamiento de la mente a través de la modelación de distintos programas que nos explicaran cómo funciona el programa en nuestro cerebro. Lo anterior tiene dos consecuencias: Primero, se logra establecer un programa de investigación que permite acceder al estudio de la mente desde la tercera persona a través de un método científico y riguroso, que logra no sólo la modelación, sino

[17] Church (en Copeland, 1993) señala algo similar, estableciendo que cualquier cálculo en el mundo real puede realizarse utilizando el "Cálculo Lambda", que es equivalente a utilizar funciones generales recursivas. Definido de manera simple, el Cálculo Lambda es un sistema formal para definir funciones y las aplicaciones de funciones, usado tanto en la Lógica Matemática como las Ciencias Computacionales.

también la experimentación con la mente. Y en segundo lugar, lograría resolver el problema mente-cuerpo toda vez que establece que el soporte físico o *hardware* en el cual la mente funciona es irrelevante (o es posible de ser realizado con distintos tipos de *hardware*), y ya sea en neuronas, cables de silicio o sistemas físicos basados en reacciones químicas, lo que importa es la organización y funcionamiento de los programas. En otras palabras, y ocupando la idea de Niveles de Descripción[18], si bien a nivel de *hardware* un computador puede ser descrito de maneras distintas - dado su soporte físico o dado que su procesador esté basado en diferentes tecnologías - a un nivel más alto podrían estar implementado el mismo algoritmo o programa, por lo que serían, en ese nivel superior, similares.

El cuarto concepto se denomina 'Test de Turing'. El Test de Turing plantea un método generado por Turing (1950), cuyo objetivo era esquivar la difícil pregunta sobre la posibilidad de que las máquinas puedan pensar: Dicha pregunta no es lo suficientemente clara sobre a qué refieren concretamente las palabras del enunciado, por lo que sería apropiado cuestionarse sobre algo más preciso y menos ambiguo. Es decir, el autor buscaba rechazar el uso común de términos tales como "máquina" o "pensar", ya que para generar consenso sobre sus significados, debería realizarse una encuesta basada sobre definiciones y frecuencias de uso de estas palabras para saber a qué refieren. Para evitar esta inconveniencia, propone el Test de Turing, que viene a ser una especie de reemplazo a la pregunta sobre si una máquina puede pensar, y lo intercambia por lo que él denomina el 'Juego de la Imitación'[19], es decir, propone preguntar si es que un computador digital puede ganar en el juego de la imitación al comparar su desempeño con el de un humano en una tarea específica.

El juego plateado por Turing tiene varias formas y matices, pero en su versión final, consiste en la interacción entre tres partes, una persona, una máquina y un juez. Cada uno de ellos se encuentra en una habitación distinta, y se les etiqueta como X e Y, indistintamente, sin que el juez sepa cuál es cuál. La tarea del juez es adivinar quién es la persona y quién es la máquina. Para tal efecto, éste debe hacer preguntas, las cuales pueden tener la siguiente forma: '¿Podría X decirme si X puede jugar ajedrez?'. El que sea X - ya sea la máquina o la persona - debe contestar. El rol en el juego para la máquina es simular que es una persona, mientras que la otra persona simplemente responde las preguntas con sinceridad. Turing se encontraba tan seguro de la eficacia del juego, que llegó a pronosticar que 50 años después de su propuesta (es decir, cerca del año 2000), la tecnología disponible (en cuanto a velocidad de procesamiento y capacidad de memoria) haría que el juez no pudiera tener más de un 70% de posibilidad de discriminar correctamente a la persona de la máquina.

En definitiva, el Test de Turing tiene la característica de ser una herramienta que permitiría obviar la discusión sobre si una máquina es inteligente o no. Dada la imposibilidad de acceder a otras "mentes" y determinar si la máquina "piensa", se apela a un criterio basado en la conducta el cual a su vez es evaluado por un juez.

[18] Los niveles de descripción corresponden al hecho de que un sistema complejo puede ser analizado y descrito de diferentes maneras a través del uso de niveles. Se considera el nivel molecular como el nivel más bajo en comparación, por ejemplo, a estructuras físicas más generales y Físicamente observables, que correspondería a la descripción más alta (baja y alta no significan necesariamente más simples o complejas).

[19] El Juego de la Imitación también puede ser definido como la actividad en la cual personas deben descubrir la identidad de uno de ellos a través de un interrogatorio.

Luego de discutir estas nociones básicas para el modelamiento de una mente por parte de la I.A., es posible presentar las arquitecturas o diseños a ser implementados, para luego ver cómo estos conceptos caracterizan la mente desde una perspectiva computacionalista.

2.2.2 Dos Arquitecturas Cognitivas.

Históricamente, la Ciencia Cognitiva ha considerado dos teorías para el diseño de una mente. Dichos diseños o 'Arquitecturas Mentales' refieren tanto a los componentes como a las relaciones y funciones entre estos últimos. El primer tipo de arquitectura se origina con el computador digital, el cual se caracteriza por el procesamiento de símbolos en un sistema físico. La segunda arquitectura, más reciente que la anterior - apareció con fuerza alrededor de 1980 - se caracteriza esencialmente por un diseño que consiste en micro-unidades profusamente interconectadas entre sí, que tienen la capacidad de activar umbrales de activación/desactivación, eliminando la necesidad de un procesador central de símbolos.

Aunque ambas aproximaciones, en última instancia, buscan reproducir conducta basada en inputs y outputs, la primera es considerada 'Top-down', ya que se encuentra basada en los principios de las ciencias computacionales, mientras que la segunda tiene a ser más bien 'Bottom-up', al poseer restricciones que reflejan propiedades neurobiológicas. Cada una, además, puede ser etiquetada con los nombres de 'Computabilidad Algorítmica' v/s 'Procesamiento de Señales', o 'Modelamiento de Inteligencia Artificial' v/s 'Modelamiento de Redes Neuronales'. A continuación, un breve resumen de cada una de estas arquitecturas.[20]

2.2.2.1 Arquitectura Clásica.

La Arquitectura Clásica del modelamiento de la mente posee ciertas características. Primero, es necesaria la presencia de sistemas periféricos que tengan la capacidad de captar los *inputs* desde el ambiente, de manera análoga a como funcionarían los sentidos en el ser humano. Segundo, existe un procesador central de información (o CPU - *C entral Processing Unit*) cuya función es ejecutar las reglas que permiten el procesamiento y transformación de símbolos, entendidos éstos como representaciones, en el que además se produce la recepción de información que entra a través de los sistemas periféricos y envía instrucciones al tercer componente de la arquitectura, los sistemas motores de *output* conductual. En la CPU se encontraría la memoria, el aprendizaje y la atención. Cuarto, la información que ingresa a la CPU se encuentra diferenciada - hecho que algunos filósofos han denominado 'Encapsulamiento de la Información' (Stillings, 1995) - lo que significa que ésta no es modificada por la CPU, sino que sólo manipulada de manera tal de generar *outputs* conductuales. En quinto lugar, y estrechamente ligado a la manipulación de la información recién mencionada, se encuentra una hipótesis básica de procesamiento para la Arquitectura Clásica denominada Hipótesis del 'Sistema Físico de Símbolos', planteada por Newell (1980). La hipótesis sugiere que la mente opera como un computador, entendiendo los fenómenos mentales como una serie de procesos que

[20] Por razones de relevancia del tema central de este trabajo, se ha dejado de lado una tercera aproximación a la actividad cognitiva, aquélla que nace como una reacción a los postulados básicos de la Arquitectura Clásica. Esta tercera aproximación es conocida como 'Cognición Corporalizada', la cual considera la cognición humana como una actividad que se encuentra estrechamente relacionada con el cuerpo humano en general, el contexto social e histórico del individuo, además de considerarla una actividad dinámica y contexto-dependiente. Según esta aproximación, ignorar aspectos tales como el contexto social, la corporalidad o constitución física del sujeto y la dependencia espacio-temporal de la actividad cognitiva *in situ* es un error, ya que entrega una visión excesivamente abstracta y lógica de la misma.

operan en función de la manipulación de símbolos, y esta manipulación de símbolos es llevada a cabo por un sistema físico real, en el caso de la mente, por el cerebro. La manipulación de estos símbolos se produce de manera sintáctica, es decir, que el uso de éstos se produce en función de su forma y no su contenido, obedeciendo al algoritmo utilizado para determinada tarea (más sobre la noción de algoritmo en 2.2.1). Finalmente, existe un concepto básico que caracteriza a la Arquitectura Cognitiva Clásica, denominado 'Representación Proposicional' (Stillings, 1995), y que consiste en sostener que las proposiciones son las unidades de pensamiento más simples, y que al relacionarse con los *inputs* sensoriales provenientes del exterior, poseen la capacidad de ser analizadas en términos de verdad y falsedad. En otras palabras, la actividad mental tiene la forma de proposiciones representacionales, y estas son concatenadas por una unidad central de procesamiento de información.

2.2.2.2 Arquitectura Conexionista.

Existe un segundo modelamiento mental, posterior a la concepción clásica de la mente. Esta nueva arquitectura deja de lado el enfoque netamente computacional con el propósito de aproximarse a la modelación de la mente utilizando una nueva metáfora: La metáfora del cerebro. Lo anterior se intenta lograr al replicar el procesamiento de información que ocurre en el sistema nervioso humano, el cual se encuentra basado en el funcionamiento de neuronas y la interconectividad entre éstas (Smolensky, 1989). Rumelhart (1989) caracteriza este modelo como una arquitectura "inspirada neuronalmente", en la cual se considera como unidad de procesamiento básico "algo así como una neurona", dejando atrás las necesidad de un procesador central encargado de la función ejecutiva de entregar instrucciones de *output* en función de los *inputs* recibidos; en este caso, es el sistema como un todo el que reacciona a los *inputs* del ambiente y produce un *output* conductual.

La utilización del funcionamiento del cerebro como metáfora trae consigo una serie de nuevos supuestos para el modelamiento de la mente. Entre éstos se encuentra el supuesto de que el conocimiento y procesamiento de información se encuentra en las conexiones entre las unidades, afirmación que implica que la actividad cognitiva se encuentra implícita en la estructura física del sistema; en otras palabras, la actividad cognitiva se encuentra "ubicada" en el funcionamiento del sistema en su totalidad (Rumelhart, 1989), y no en algún componente central que procesa información, ni tampoco se encuentra explícitamente dada en la forma de proposiciones como la Arquitectura Clásica sugiere. En el modelamiento clásico, la información se encontraba en los estados de los elementos. Sin embargo, si bien puede existir algún tipo de memoria de trabajo que permita almacenar información de corto plazo en el modelamiento conexionista, la información de largo plazo se encuentra entre las conexiones de las microunidades.

Adicionalmente, se produce otra diferencia con la arquitectura anterior, pero a un nivel más abstracto. Esta diferencia radica en la gran cantidad de variables y limitaciones a las cuales la conducta humana se encuentra sometida, y una Arquitectura Conexionista permitiría reaccionar de mejor manera a estas últimas; la Arquitectura Clásica contempla el funcionamiento de algoritmos que funcionan de manera serial dada su naturaleza de diseño, es decir, se produce una computación determinada y luego la siguiente. No obstante, los algoritmos conexionistas tienen la propiedad de operar en paralelo, hecho que la asemeja a la manera en la cual el mismo cerebro funciona. Lo anterior permite establecer una manera distinta de conceptualizar la conducta, ya que ésta comienza a considerarse no como el producto de un sólo componente del sistema cognitivo, sino como producto del

funcionamiento de múltiples componentes[21], que en un trabajo simultáneo y de mutua interacción, genera el funcionamiento del sistema y en definitiva, respuestas de *output*.

En conclusión, y más allá de las dos arquitecturas recién descritas, para comprender las propuestas de la Ciencia Cognitiva es esencial una concepción de vida mental en términos computacionales, tanto desde un punto de vista metodológico así como teórico. Metodológico, ya que permite un acceso objetivo y externo a su funcionamiento lo que a su vez permite su modelación, y teórico como hipótesis central para la "construcción" de una mente. Por lo tanto, se asume que ciertas características de un objeto real del mundo (la mente) comparte ciertos rasgos con un sistema idealizado e hipotético (un computador y su funcionamiento), entregándole un mejor fundamento a la metáfora del computador como modelo cognitivo, lo que permite considerar a la mente como un mecanismo computacional virtual.

2.3 Cuando Metáfora y Realidad se Confunden: Inteligencia Artificial Fuerte.

La visión funcionalista-computacionalista clásica de la mente que se ha descrito puede sintetizarse en un par de oraciones: La manera en la cual funciona nuestro sistema cognitivo es que el cerebro es un computador digital, y la mente correspondería al *software* que opera en este computador, siendo los estados mentales, por lo tanto, estados computacionales del cerebro. Esta visión computacionalista de la mente también es conocida con el nombre de 'Inteligencia Artificial Fuerte', etiqueta otorgada por John Searle en su célebre artículo 'Minds, Brains and Programs' (1980). Según el autor, en Inteligencia Artificial existen dos posturas aglutinantes que lidian precisamente con el rol del computador en la comprensión de la mente: La primera, catalogada como 'Inteligencia Artificial Débil', considera al computador, máquina o programa como una "herramienta" que nos permite acceder a los procesos mentales a través de la simulación/implementación de estos últimos. Por otro lado, la Inteligencia Artificial Fuerte plantea que al modelar una mente se está, efectivamente, creando una nueva mente; es decir, un computador que modela un proceso mental no es un modelo, sino más bien, una mente propiamente tal. Por lo tanto, es posible decir que la Inteligencia Artificial Débil considera al computador como una metáfora de la mente, mientras que la Inteligencia Artificial Fuerte considera al computador literalmente como una mente, es decir, un computador posee estados mentales, al igual que los que tendría un ser humano al momento de comprender una historia o relato[22]; en otras palabras, estamos en la presencia de la distinción modelación v/s realización, respectivamente.

[21] Lo anterior se opone directamente a teorías posteriores que buscan modelar la mente sin incluir la idea de un procesador central que lo controla todo. Entre éstas se encuentran la 'Cognición Situada', aproximación que enfatiza el carácter situado de la actividad cognitiva, es decir, ésta opera en medio ambientes que estructuran, dirigen y moldean los procesos cognitivos. La cognición, por tanto, sería un proceso dinámico de interacción entre cuerpo y ambiente, que opera en un constante intercambio de información dando forma a los *outputs* conductuales, sin necesariamente requerir un procesador central que se encargue de la actividad cognitiva. Es decir, la cognición se produce *durante* la interacción entre agente y ambiente.

[22] La referencia a la comprensión de historias refiere al trabajo de Schank y Abelson mencionado el propio artículo de Searle. El programa de Schank y Abelson buscaba replicar la habilidad humana para comprender historias, basado en el hecho de que la capacidad de comprensión humana permite responder preguntas sobre ésta incluso en casos en los que la información nunca se diga explícitamente.

Con el advenimiento de esta nueva propuesta, se podía prever una solución definitiva al problema mente-cuerpo, toda vez que eliminaba el halo de misterio que rodeaba a la conexión entre nuestra dimensión material y mental, y la reemplazaba por una conexión exenta de dicho problema. La solución consistía en sustituir la relación mente-cuerpo por una entre programa y computador.

2.3.1 Resumiendo: La Mente según la I.A.Fuerte.

Según el Funcionalismo, existen dos maneras de clasificar los objetos en el mundo: Aquellos que se caracterizan por su estructura física (ejemplo: Una manzana, una célula, el vino) y aquellos que se caracterizan por su función (ejemplo: Un carburador, un escenario, una mesa). Los primeros deben poseer una determinada e intrincada estructura física que los hace ser lo que son, y se les identifica por esa composición porque, de manera adicional, ésta determinará cómo se relacionan con el resto de las leyes de la naturaleza. Los segundos pueden estar constituidos por una variada gama de componentes y estructuras físicas, y se les identifica por sus actividades o tendencias a funcionar de determinada manera; por ejemplo, un escenario, sin importar de qué está construido, se utiliza para exhibir algo destinado a ser visto por una audiencia, por lo que es caracterizado por su función, mas no por su estructura física.

Por lo tanto, la pregunta surge: ¿Dentro de cuál de las dos maneras de clasificar el mundo ubicamos la actividad mental? Es decir, aquello que denominamos como mente, ¿Se caracteriza por su estructura física o por su función? Según el Funcionalismo, sería un error asumir lo primero. Sin embargo, de acuerdo con ciertas corrientes reduccionistas, la mente correspondería a la actividad cerebral. Por lo tanto, asumiendo que el cerebro se ubica entre las clases físicas del mundo, es natural especular sobre la posibilidad de que las mentes sean cerebros, lo que haría que nuestras mentes también sean físicas. Una visión así concede preponderancia a disciplinas tales como la Neurociencia para entender la cognición, ya que al entender la constitución física del cerebro, en realidad se están estudiando los estados mentales.

No obstante, diría un funcionalista, cabe la posibilidad de que la actividad mental se desarrolle utilizando un tipo de material físico distinto, lo que implicaría que la mente debiera ser clasificada según su función dentro del sistema cognitivo, y no en función de su constitución física. Por ejemplo, imaginemos un mundo posible en el que todas las mesas fueron y son, históricamente, sólo de madera. Sería difícil pensar que pudiera haber una mesa de acero, porque dada nuestra experiencia e intuición, una mesa debería ser de madera para que sea una mesa. Entonces, cabe preguntarse por la posibilidad de que, quizás, una mesa siga siendo una mesa incluso si ésta está hecha en función de otro material. Según el Funcionalismo, lo mismo ocurre con la actividad mental. Quizás sí sea posible que las mentes puedan ocurrir en materiales distintos al cerebro. Por lo tanto, el Funcionalismo plantea que algunas mentes podrían ser cerebros, siempre y cuando posean una organización tal que les permita procesar información, controlar la conducta y realizar actividades mentales tales como almacenar información, incorporar nuevos conocimientos, así como otras tareas asociadas a la cognición. Es decir, la mente sería una propiedad funcional, no una física.

Es sobre esto último sobre lo cual la I.A. basa sus ideas y el modelamiento de máquinas, ya que libera a la mente de una realización específica para poder utilizar otro tipo de *hardware* el cual también podría exhibir capacidades que parecieran requerir algún tipo de actividad mental similar a la humana, y permite la posibilidad de replicar actividad cognitiva

en criaturas diseñadas por el hombre en función de otro sustrato físico. La influencia de la idea del Funcionalismo de diferenciar los estados mentales de la materia que los origina, sumado al hecho de caracterizar a los estados mentales como estados funcionales, da pie a la posibilidad de implementar estos sistemas en programas y/o máquinas para simularlos computacionalmente, y así entender mejor la manera en la cual los fenómenos mentales operan, de nuevo, no apelando a ningún tipo de reducción ni eliminación de los estados mentales, sino por el contrario, viéndolos funcionar en tareas concretas, o bien comparándolos con conductas humanas ante tareas similares.

Hay un argumento fundamental considerado por la I.A., el cual proviene de una discusión filosófica entre la Teoría de la Identidad (sección 1.2.3.2) y el Funcionalismo. El argumento tiene la forma de un experimento mental. Imagine un *gedankenexperiment* en el cual se descubre vida en Marte, y que no sólo eso, éste se encuentra poblada por marcianos. También se descubre que dichos marcianos tienen una determinada constitución fisiológica (para propósitos del experimento mental, dicha constitución se basa en células marcianas que funcionan con hierro en vez de las células terrestres que funcionan con carbón). El cerebro de dicho marciano, por lo tanto, tendría una constitución y ensamblaje distinto al del ser humano. No obstante, el funcionamiento de ese cerebro podría exhibir una economía cognitiva funcional de estados internos similar a la humana, una economía funcional en la cual exista una relación entre los estados del cerebro humano y los del marciano. Es decir, podríamos identificar y etiquetar estados mentales tales como 'Dolor' en el cerebro marciano sólo en función de las relaciones de ese estado mental con otros estados mentales que a su vez se manifestarían en su conducta. Ahora bien, la crítica a la Teoría de la Identidad radica en el hecho de que ese estado mental identificado sería totalmente distinto al de un humano desde un punto de vista fisiológico, aunque éste seguiría siendo idéntico desde un punto de vista funcional, lo que sería una crítica directa a la idea de la Teoría de la Identidad de que los estados mentales no son más que estados cerebrales.

Y el argumento no se detiene ahí. Además, establece que si la economía interna de los estados mentales del marciano fuera isomórfica funcionalmente con la economía interna de estados mentales humana (en otras palabras, si esos estados estuvieran conectados causalmente a *inputs* externos, conectados a otros estados mentales y a la conducta de manera paralela a nuestras propias conexiones), entonces nuestro marciano podría sufrir dolor, tener deseos, creencias y miedos al igual que un ser humano, sin importar las diferencias de sustrato físico entre ambos organismos. El comentario que el Funcionalismo destaca y quiere comprobar es que lo importante para la vida mental de una criatura no es la materia con la cual la criatura está compuesta, sino que la estructuración de las actividades internas de esa vida mental.

Es desde esta idea que la I.A. encuentra cabida para la creación de criaturas basadas en algún otro compuesto distinto a la materia orgánica humana. Es decir, si es posible concebir la posibilidad de otras constituciones físicas distintas a la humana - en este caso, la de un marciano - es posible plantear la existencia de un dispositivo artificial electrónico o algún computador que pueda poseer estados mentales si es que éste es dotado de una economía interna que sea funcionalmente isomórfica a la nuestra.

El Funcionalismo, en conclusión, plantea que es lógicamente posible que existan seres con mentes como las nuestras, con nuestros estados mentales, que no tengan cerebros y sistemas nerviosos basados en carbono. Lo anterior significa un complicado problema a la Teoría de la Identidad, ya que, al parecer, podría haber más de un estado físico al cual un estado mental podría corresponder. Lo anterior, simultáneamente, abre la puerta para la

creación de criaturas diseñadas por el hombre que sean capaces de exhibir capacidades mentales humanas. En otras palabras, la propuesta funcionalista plantea que, dado que un estado mental puede tener múltiples realizaciones materiales, existen maneras artificiales de crear mentes y estados mentales.

La I.A. Fuerte se basa en la concepción funcionalista de la mente para el diseño y modelación de actividad mental; ésta es adecuada debido a que permite liberar a la Ciencia Cognitiva de la necesidad de comprender cómo funciona el cerebro y cómo se relaciona con el cuerpo humano con la mente, y si bien esta última puede estar asociada a ciertas estructuras físicas para que funcione, éstas son irrelevantes a la estructura profunda de la actividad mental. Lo anterior permite abstraer la mente del cuerpo, permitiendo a la I.A. Fuerte abocarse a la comprensión del funcionamiento de nuestra vida psicológica en términos de cómo opera un computador, mediante descomposición recursiva de su funcionamiento.

Volviendo a los conceptos anteriormente presentados en la sección 2.2.1, y a manera de recapitulación de esta sección, la I.A. Fuerte sostiene que la mente es un dispositivo funcional situado en un medio ambiente cambiante. El medio ambiente entrega *inputs* al dispositivo mental, cambiando sus estados internos, los que a su vez se ven reflejados en su *output* conductual. Lo anterior se produce ya que el dispositivo mental realiza un manejo de información a través de la manipulación de representaciones provenientes del medio ambiente los cuales, últimamente, se ven reducidos a la manipulación de símbolos[23]. El cerebro sería un computador digital, probablemente, una Máquina Universal de Turing, la que lleva a cabo algoritmos a través de la implementación de programas, y aquello que llamamos mente, es un programa o un grupo de programas, cuyo procesador de información utiliza componentes muy básicos y reglas de transformación sencillas para llevar a cabo una tarea determinada, sin necesidad de nada adicional para desempeñar dicha tarea.Con respecto a los estados mentales tipo (tales como creer, desear, esperar, etc.), éstos también serían caracterizados de manera funcional; es decir, éstos serían estados mentales que tendrían relaciones de causa-efecto con *inputs* y otros estados mentales, y la actividad mental se explicaría por reglas sintácticas, programas, que debemos descubrir, los cuales otorgan la llave para entender la mente y la inteligencia.

La tarea de la Ciencia Cognitiva para comprender las capacidades mentales consistiría en descifrar cuáles son los programas implementados cuando dichas capacidades operan, dejando de lado la tarea de encontrar relaciones *type-type* con el cerebro, debido a la posibilidad de utilizar distintos tipos de *hardware* para el desarrollo de la actividad mental. La cognición se encuentra reducida a la actividad simple de ceros y unos (u otros símbolos), por lo que la manera de determinar si determinado programa implementa efectivamente capacidades cognitivas será a través de mediciones de la conducta del programa (Test de Turing), lo que a su vez permite a los psicólogos realizar mediciones y comparaciones entre los programas utilizados por las máquinas y los utilizados por los humanos.

La visión computacionalista de la mente sostiene que la mente es un computador digital, un dispositivo de estados finitos que almacena representaciones simbólicas y las manipula a través de reglas sintácticas. Los pensamientos serían representaciones mentales simbólicas, y los procesos mentales serían secuencias causales dirigidas por las características sintácticas (y no las semánticas) de aquellos símbolos. Es a esta visión

[23] En el sentido abstracto de la definición de computabilidad sería posible decir que hígados, estómagos, cerebros y plantas computan. Lo que algunos filósofos destacan como diferencia entre el cerebro y esos ejemplos sería la capacidad de éste para representar el mundo exterior, y a través de la noción de computación, y de producir conducta motora en tiempo real.

funcionalista de la mente a la cual en el próximo apartado se dirigen las principales críticas y cuestionamientos.

2.3.2 Un Problema Insalvable: Conciencia v/s Funcionalismo.

Hasta este punto, se ha descrito la visión dominante en Filosofía de la Mente, al menos desde 1970. Y si bien trajo consigo respuestas y planteamientos relativos a una nueva perspectiva para concebir la mente (además de un programa de investigación), también es cierto que presenta una gran variedad de objeciones, particularmente relacionado con el hecho de que el Funcionalismo deja de lado ciertos aspectos de la mente que forman parte de nuestra realidad cognitiva. A continuación se citarán algunas observaciones sobre los conceptos utilizados por el Funcionalismo, para luego concentrar críticas sobre el problema de la conciencia.

Si se analiza su definición con precisión, el argumento a favor del Funcionalismo comienza con la premisa de que es “posible” que una mente pueda ser implementada en algún sustrato físico distinto al cerebro. Sin embargo, ¿Qué justifica la aseveración de que algo así pueda ocurrir? Si lo posible es lo imaginable, también puedo imaginar que sería posible que fuera imposible que la mente se dé sin cerebro. Es decir, el hecho de que sea posible imaginar una mente sin cerebro no es razón suficiente para sostener que esto pueda ocurrir empíricamente. Lo que en realidad hace falta es una manera razonable de justificar premisas sobre lo que es posible o lo que podría ser considerado como posible con respecto a este problema. Hasta que el Funcionalismo pueda certificar la posibilidad de una mente fuera de un cerebro, el argumento parece inconcluso, mas no inválido.

Una segunda crítica se encuentra asociada a la utilización de un criterio conductista para determinar la presencia de actividad cognitiva, i.e., comentarios sobre la utilización del Test de Turing como una manera óptima de validar la replicación de actividad cognitiva. Es decir, ¿Por qué la duplicación de la conducta lingüística por parte de una máquina programada parece evidencia suficiente para considerar que éstas piensan y/o son inteligentes? Es decir, vale la pena preguntarse por qué Turing considera que un juez pueda ser engañado por el comportamiento lingüístico de una máquina como razón suficiente para establecer que ésta sea efectivamente inteligente. Es sensato suponer que Turing busca posicionar un argumento epistemológico sobre cómo saber si una máquina posee inteligencia, pero aquello es distinto al hecho concreto de que la máquina sea inteligente. Podría darse el caso que, efectivamente, la máquina sea inteligente sin que ésta convenza a un juez de que lo sea, o que no exista un juez que lo evalúe. En otras palabras, que la máquina sea inteligente no depende de un juez o el convencimiento de éste para que la máquina posea tal propiedad. Lo anterior proviene del planteamiento de Turing sobre el Juego de la Imitación antes presentado. Sin embargo, es plausible cuestionarse: ¿Es la imitación de una propiedad humana en términos conductuales suficiente para establecer que esa máquina replica propiedades como la inteligencia o la conciencia? Es posible sostener entonces que el Juego de la Imitación posee un sesgo conductista, que produce reducción del resultado conductual de la actividad cognitiva sobre la conducta cognitiva misma. Lo anterior resulta contradictorio, ya que la inteligencia no necesariamente se encuentra en el ojo del observador, toda vez que ésta posee un rasgo esencial que la diferencia de otros fenómenos de la ciencia: La actividad cognitiva le ocurre a un individuo, a un punto de vista, y mucha de esa actividad cognitiva podría no verse proyectada en dicha actividad conductual (o al menos no verse proyectada de manera íntegra), por lo que

reducir inteligencia a conducta podría constituir un error[24]. Y si bien el Test de Turing permite superar el problema plateando por el autor en relación con la dificultad inicial de definir "Inteligencia" y "Pensar", su solución no es efectiva, ya que no es posible decir que el Test demuestre de manera definitiva que una máquina posea dicha característica. Es posible concluir, por lo tanto, que la calidad de la simulación de una propiedad mental no implica la presencia del objeto simulado (Searle, 1980), y que el Test de Turing asume erróneamente que es posible la simulación de inteligencia a través de un mecanismo artificial que simule su conducta. No es posible reducir un método evaluativo de inteligencia a aquello que produce la actividad mental, ya que los factores causales de ésta podrían perfectamente ser independientes de los jueces, produciendo una distancia entre los aspectos evaluativos o epistemológicos de la conducta con el aspecto ontológico de la actividad cognitiva, lo que es un problema para la idea inicial de Turing de zanjar el problema de las otras mentes y evitar la discusión sobre qué es pensar y qué es inteligencia.

Por otro lado, al parecer el Funcionalismo también tiene problemas con respecto al contenido de las representaciones mentales tal cual las concibe - como entes netamente funcionales. Los estados mentales son estados representacionales con un contenido semántico. Por ejemplo, mientras Ud. juega ajedrez, está pensando sobre el juego. Se da cuenta que en las próximas jugadas podría perder a la reina, pero que en realidad forma parte de una trampa que le quiere tender a su contendor. Se da cuenta de los riesgos, sin embargo, decide continuar con la estrategia, en algo así como un arranque de "valentía ajedrecística", valentía que dicho sea de paso, no se encuentra expresada en ninguna regla del juego. Ahora, considere la creación de un computador que ha sido programado para jugar emulando perfectamente la manera en la que Ud. juega, algo así como una réplica funcional de su manera de jugar ajedrez. Para el Funcionalismo, tanto el computador como Ud. exhiben los mismos estados mentales. No obstante, ¿El computador realmente piensa de la misma manera en la cual Ud. piensa al jugar ajedrez? Intuitivamente, se podría sostener que el computador no se da cuenta genuinamente, por ejemplo, de que hay un riesgo que podría hacerlo perder la partida si es que la trampa con la reina no resulta, ya que el computador se encuentra imitando su conducta y sus estados representacionales a través de un método muy distinto al cual Ud. representa y navega el mundo usando sus estados mentales. Esto significa que si el computador y Ud. difieren en la manera en la cual representan el mundo, o si Ud. representa el mundo apelando tanto a un contenido semántico (dado por la manera en la cual los humanos representamos el mundo) así como sintáctico (entregado por las reglas de funcionamiento del ajedrez y distintas estrategias de juego) mientras que el computador sólo apela a un componente sintáctico, es razonable suponer que quizás el Funcionalismo haya dejado de lado un aspecto fundamental de nuestra cognición. Por ejemplo, al computador le sería imposible tomar una decisión basada en "valentía ajedrecística", ya que ésta no se encuentra en ninguna regla del juego del ajedrez, y si bien se podría argumentar que una variable de juego aleatorio podría ser incorporada al momento de crear un programa que imite mi juego o que el concepto de "valentía ajedrecística" podría ser definido en términos de reglas y representaciones, ésta

[24] Putnam (1968) realiza una interesante crítica al Conductismo Lógico, estableciendo que sus postulados no solucionan el problema planteado por Descartes. Uno de los argumentos que el autor utiliza es un experimento mental en el cual se plantea la existencia de un mundo hipotético, en el cual las personas sienten dolor pero no lo expresan en su conducta. Es decir, una persona que se encuentra experimentando algún dolor se comporta exactamente igual que una persona que no lo tiene. Según las ideas del Conductismo Lógico, se debiera concluir que estas personas no tienen ese dolor. Pero Putnam replica incluyendo la posibilidad de medir de alguna manera las ondas de dolor producidas por el cerebro, así demostrando que los habitantes de este mundo hipotético se encuentran, de hecho, sufriendo, situación que presenta problemas para la relación pensamientos-conducta.

carece del valor semántico[25] que representa para mi esquema de juego, ya que este súbito arranque de coraje y riesgo no se encuentra "pre-cargado" en mi sistema de juego.

Hasta acá, se han mencionado observaciones sobre algunos de los conceptos planteados por el Funcionalismo para sostener una teoría funcionalista de la mente. Teoría, que como se ha mencionado, ha sido recogida por la I.A. Fuerte para la modelación de máquinas que exhiban capacidades cognitivas. Sin embargo, el Funcionalismo falla en dar cuenta de una de las principales características de la cognición humana, aquella relacionada con la conciencia. Por el momento, definiremos la conciencia como aquella dimensión de la actividad cognitiva que es subjetiva e íntima al momento de experimentar el mundo. El problema con el Funcionalismo es que no puede explicar el hecho de que, por ejemplo, dos personas puedan diferir en la manera en la cual aprecian un vino. El sujeto Luis y el sujeto Cristóbal pueden ser funcionalmente idénticos al momento de catar un vino, pero también pueden tener una experiencia subjetiva distinta al momento de beberlo. Ambos podrían, por ejemplo, haber asistido al mismo curso de cata de vinos, haber obtenido la misma evaluación en dicho curso, poseer características cognitivas normales y elegir a la misma cepa como la mejor cepa que han tomado; sin embargo, la experiencia subjetiva o los *qualia* de beberlo podrían ser diferentes. Y no hay nada en la concepción algorítmica de la mente que lo explique ni lo prevea. Es decir, dos personas que sean funcionalmente idénticas pueden diferir en la manera en la cual experimentan el mundo, ya que lo que se considera como congruencias a nivel funcional pueden enmascarar diferencias sustanciales a nivel personal, es decir, es posible que hayan diferencias en la experiencia y cognición de dos sujetos sin que éstas se manifiesten ni en lo funcional ni lo conductual[26]. En definitiva, si las experiencias cualitativas distintas diferencian los estados mentales de nuestra vida mental, una explicación funcionalista no da cuenta de éstas.

En una línea argumentativa similar, Chalmers (1996) presenta el problema de los '*Qualia* sensoriales'. Según este filósofo, los *qualia* se definen como cualidades subjetivas de las experiencias mentales. Cuando una persona se encuentra en un estado mental consciente, existe algo así como la experiencia subjetiva de tener ese estado. Piense, por ejemplo, en la sensación que le produce pasar un tenedor afilado por sobre el fondo de una olla, o la sensación de escuchar una canción de *heavy metal*. Esa sensación privada, subjetiva y personal se le denomina *quale*[27], y cada estado mental consciente es un *quale*, debido a que cada estado mental consciente se siente de determinada manera. Apelando a otro tipo de definición, los *qualia* también pueden ser identificados como lo doloroso del dolor, o lo placentero del placer.

Al poseer esta característica subjetiva y personal, las propiedades de la experiencia sensorial no pueden ser definidas ni descritas objetivamente, ya que sólo pueden ser experimentadas directamente, desde la primera persona, lo que además da paso a un gran problema epistemológico con respecto a cómo definirlos y estudiarlos. Es decir, los *qualia* sensoriales sólo pueden ser aprehendidos cuando se les experimenta de manera personal, directa, y no a través de una observación externa desde la tercera persona.El problema en concreto del Funcionalismo con los *qualia* sensoriales es el siguiente: Al presentar a los estados mentales como propiedades esencialmente relacionales, este ignora la naturaleza cualitativa de cualquier estado mental. La teoría funcionalista no considera esta propiedad cualitativa en su descripción, o mejor dicho, es incapaz de dar cuenta de ésta; no puede

[25] Ese valor semántico se relaciona con la Intencionalidad que hay detrás. Más sobre este punto en el Apéndice.

[26] Un clásico ejemplo es el experimento mental del 'Espectro Invertido', que se desarrollará más adelante.

[27] *Quale* corresponde al singular del plural *Qualia*.

explicar por qué un estado mental posee determinada experiencia subjetiva, ni tampoco puede definir funcionalmente dicho aspecto cualitativo.

Más aún, sería posible expandir la idea y contemplar la posibilidad de que existan diferencias entre las experiencias subjetivas de cada individuo, para las cuales el Funcionalismo no ofrece explicación. En la literatura existen varias versiones de lo que se denomina problema del 'Espectro Invertido' (formulado por primera vez en Locke, 1690). El Argumento del Espectro Invertido consiste en lo siguiente: Imagine que Pedro y Pablo son dos seres humanos cuyas capacidades físico-psicológicas son absolutamente normales, por lo que son capaces de realizar las mismas discriminaciones somato-sensoriales que cualquier individuo haría, ambos exhiben plena y completa normalidad con respecto a las funciones de su cuerpo, por lo que, por ejemplo, pueden realizar los mismos tipos de discriminaciones de color. Si se les pide a los sujetos que tomen la manzana roja y dejen la manzana verde dentro de una cesta, ambos elegirán correctamente la manzana roja, o si se les pidiera que tomen una manzana verde y se la arrojen a su oponente, lo harían sin ninguna dificultad. Pero supongamos que la experiencia subjetiva de cada uno de ellos es absolutamente opuesta. La experiencia que Pedro denomina como "veo rojo" corresponde a la experiencia que Pablo denomina "veo verde", y viceversa. En este caso, se produce lo que se denomina una "experiencia invertida de rojo-verde" (Figura 1).

Figura 1.

Si el espectro invertido fuera posible, entonces el Funcionalismo sería incapaz de explicar las experiencias subjetivas, ya que éstas simplemente se encuentran fuera del alcance de esta aproximación, lo que generaría un error en la descripción de los estados mentales: El Funcionalismo entregaría exactamente la misma descripción de la experiencia subjetiva de Pablo cuando dice "veo rojo" que la experiencia subjetiva de Pedro cuando dice "veo rojo", siendo que ambas experiencia son distintas; el Funcionalismo establece que los estados mentales de Pedro y Pablo son funcionalmente isomórficos, haciendo suponer que si un estado mental tiene la función de ser "la sensación de rojo", entonces por definición es un estado de rojo. En otras palabras, una aproximación funcionalista a los estados fenoménicos de Pedro y Pablo sería errónea, ya que desde el punto de vista conductual-funcional, éste diría que los estados de ambos sujetos son iguales, siendo que no lo son.

Tanto Pedro como Pablo adquirieron los términos rojo y verde a través de sus experiencias personales, subjetivas y privadas, y por eso no es posible discriminar si, efectivamente, tienen experiencias diferentes desde el punto de vista funcional y externo. Por lo tanto, El Funcionalismo fallaría al momento de explicar dichas experiencias cualitativas.

El Espectro Invertido es totalmente concebible, no sólo desde un punto de vista filosófico; también es posible encontrarlo en casos clínicos verídicos, como el Daltonismo[28] o lo que se conoce en la literatura como 'Visión Pseudonormal'[29] (Nida-Rümelin, 1999). El Funcionalismo establece que el Espectro Invertido es imposible ya que en su definición de estados mentales no hay espacio para los *qualia*. Si a eso se suma el hecho de que el Espectro Invertido es posible tanto filosóficamente como clínicamente, por lo tanto, es razonable suponer que el Funcionalismo es falso (Churchland, 1988), o al menos, incorrecto con respecto a este punto.

Algún filosofo funcionalista podría contra-argumentar que la presencia de la experiencia subjetiva no necesariamente representa un problema para su posición, toda vez que la teoría puede aceptar que diferentes individuos presenten *qualia* diferentes, sin afectar el hecho que los estados mentales son esencialmente estados relacionales; según este filósofo, la explicación del estado mental que ellos proporcionan puede prescindir de los *qualia*. No obstante, una respuesta como la que entregaría este filósofo otorga la oportunidad de abrir la ventana a la idea de que exista algo así como la "Ausencia de *Qualia"*, es decir, un sistema funcionalmente idéntico a nuestro organismo pero que no presente las experiencias subjetivas que nosotros exhibimos. Pero lamentablemente, eso incurriría en el error de ignorar el que los *qualia*, por redundante que suene, existen, son reales, y son una capacidad cognitiva a la cual todos podemos acceder. El hecho de no poseer las herramientas para poder dar cuenta de tal capacidad no es motivo suficiente para caer en un reduccionismo o un eliminativismo. El problema con la identificación de estados mentales con estados neuronales es que existe la posibilidad de que se presente unos sin los otros. Y si bien el reduccionismo admite la existencia de los *qualia*, afirma que éstos no serían más que estados mentales X.

Por otro lado, el argumento de la inversión rojo-verde tiene una consecuencia colateral. Esta consecuencia permite generar una crítica a cualquier método de estudio que incluya una observación puramente externa de los fenómenos mentales. Es posible concluir que ningún tipo de prueba conductista nos permitiría detectar la experiencia invertida de rojo-verde, porque la finalidad de una prueba como ésta es detectar la capacidad de realizar discriminaciones con objetos del mundo, y no la capacidad de etiquetar experiencias subjetivas internas. El problema entonces radica en que Pedro y Pablo pueden tener experiencias subjetivas del mundo muy distintas, pero el comportamiento de ellos seguiría

[28] El Daltonismo es un defecto genético que consiste en la imposibilidad de distinguir algunos colores, también conocida como*discromatopsia*.

[29] Normalmente, se considera la inversión de *Qualia* en experimentos mentales. Pero existe evidencia científica basada en la experiencia psicológica de la visión en colores y teorías sobre la genética de las deficiencias en visión en colores que llevan a la predicción de una condición denominada 'Visión Pseudonormal', la que correspondería a la inversión de la experiencia subjetiva de ver rojo y verde. Lo anterior se produciría debido a que los fotopigmentos contenidos en los conos que juegan un rol central en la visión humana en colores de rojo y verde se encontrarían intercambiados debido a un problema genético. Es decir, el fotopigmento sensible para las ondas de color rojo se encontraría dentro de los conos destinados para percibir verde, y los fotopigmentos sensibles a las ondas para el color verde se encontrarían dentro de los conos destinados a percibir rojo. Debido a esto, las capacidades discriminatorias de una persona con esta condición se mantienen intactas, sólo que cuando una persona Pseudonormal ve un tomate, lo ve con el *quale* asociado a verde de una persona normal.

siendo exactamente el mismo, lo que invalida la presunción de que los estados mentales pueden ser estudiados puramente empleando métodos objetivos basados en la conducta.

En conclusión, las críticas que se pueden formular al Funcionalismo consisten básicamente en destacar el cuestionamiento a una teoría que insiste en negar realidades que provienen del sentido común y nuestra vida cotidiana, es decir, ignorar nuestros estados conscientes o nuestra capacidad de sentir, o el hecho de que "algo nos pasa" cuando tenemos experiencias subjetivas sensoriales tales como comer, o sentir miedo. Por sobre todo, lo criticable - e incluso sospechoso - es que ciertas teorías insistan en negar la existencia de las experiencias conscientes como un hecho científico real, tan real como las secreciones pancreáticas o la caspa, es pos de sostener determinada teoría (más sobre este punto en sección 4).

No obstante, la conciencia es un tema controversial en Filosofía de la Mente. Algunos autores sostienen que no existe claridad respecto de determinar sobre qué se habla cuando se habla de conciencia, lo que hace a la idea poco concreta, y que en realidad, es un tema menor, poco relevante tanto para el Funcionalismo como para el diseño de máquinas inteligentes. Es por eso que se hace necesario discutir sobre qué hablamos cuando hablamos de conciencia, de manera tal de identificarla y ver cómo ésta se relaciona con el Funcionalismo, la I.A. Fuerte y el interés que ésta debiera mostrar por replicar esta capacidad cognitiva. Justamente, este tópico será abordado en el siguiente capítulo.

3. El Problema: La Conciencia.

3.1 ¿De qué Estamos Hablando Cuando Hablamos de Conciencia?

La Ciencia Cognitiva ha intentado dar cuenta de todo aquello que acaece en la mente humana. Desde sus orígenes en el verano de 1956 en la conferencia de Darmouth, esa ha sido la empresa fundamental de la disciplina. En la actualidad, y como se ha señalado anteriormente, la visión funcionalista de la mente ha sido el paradigma que ha adquirido mayor popularidad. No obstante, hay un aspecto sobre la mente para el cual el computacionalismo no ha podido entregar una explicación satisfactoria: La conciencia. El gran desafío es cómo dar cuenta de nosotros, "seres conscientes, poseedores de una mente, libres, racionales, dotados de lenguaje, sociales, además de ser agentes políticos" (Searle, 2004, p. 7) en un mundo en el cual la ciencia sostiene que todo es material y físico, por lo que todo debe ser entendido y explicado en esos términos.

Las experiencias conscientes tienen la particular característica de ser, probablemente, el aspecto mental más cercano a nosotros, el más directo y accesible. Sin embargo, parece ser el más escurridizo en cuanto a su explicación. Las experiencias subjetivas de las cuales todos somos objeto son un hecho concreto, pero hasta el momento no ha sido posible saber cómo ni por qué se producen. Sin embargo, de todas las aproximaciones al problema de la conciencia, son las que reducen filosóficamente los estados conscientes a lo material o aquellas que los eliminan las que gozan de mayor aceptación.

Inmediatamente, cabe cuestionarse cuál es la razón para adoptar una posición que busca la reducción/eliminación del problema. Las respuestas son variadas. Primeramente, es posible establecer que la conciencia posee un estatus problemático con respecto a su definición, así como su *deixis* (siendo esta última considerada como una posible solución a la primera). La problemática radica en el hecho de que la 'Conciencia' - entendida como la experiencia subjetiva del mundo que un individuo siente - es un fenómeno al cual todos podemos acceder de manera personal y privada, y todos podemos entender a qué nos referimos cuando hablamos de conciencia; no obstante, si bien entendemos a qué referimos, al mismo tiempo es difícil estudiarla de manera objetiva, es decir, apelando a la observación de tercera persona. Es precisamente esa inefabilidad la que dificulta su estudio, y no sólo eso, es la que produce un problema sin solución al momento de determinar cuál es la mejor manera de tratarla y aproximarse metodológicamente a la misma desde la tercera persona. Baars (1993) refiere a este problema de la siguiente manera:

[La experiencia subjetiva es] el bulto más confuso y discutible de la ciencia psicológica. Todos somos seres conscientes, pero la conciencia no es algo que podamos observar directamente, sólo se puede hacer dentro de nosotros mismos, y por lo tanto se hace únicamente en retrospectiva (p. xv, traducción mía.).

La conciencia, por tanto, se convierte en el problema más difícil de circunscribir en las ciencias de la mente. Lo anterior se debe a que no hay nada más íntimo y real que nuestra

experiencia consciente, pero nada más difícil de explicar, de transmitir, de exhibir a otros, es decir, de abordar objetivamente, desde la tercera persona. Es que al parecer, mientras más nos acercamos a la descripción objetiva, más nos alejamos de esa subjetividad que la hace tan particular. La conciencia fenoménica posee un aspecto cualitativo el cual se pierde al tratar de estudiarla sin considerarlo. Lo anterior se debe a su carácter personal, es decir, la conciencia le ocurre a un sujeto, tiene una ontología de primera persona, y los métodos de las Ciencias Naturales emplean métodos cuya ontología es de tercera persona. Piense, por ejemplo, en las dificultades que implica medir la experiencia subjetiva de escuchar música a diferencia del estudio de los mecanismos físicos que se desencadenan en el oído al hacerlo, o bien, compárelo con el estudio de la medición de conductas asociadas a tareas inteligentes que miden velocidad o discriminación de colores. La actividad cognitiva normalmente presenta una conducta asociada a ésta, y este desempeño puede ser medido por algún tipo de prueba. Pero la experiencia subjetiva puede existir sin que haya dicha conducta, como cuando nos encontramos en nuestra cama sin recibir ningún tipo de estimulo ni manifestar algún tipo de conducta, sin embargo, nos encontramos en un estado consciente el cual posee un componente cualitativo concreto, real.

Han existido muchos intentos por dar cuenta de la conciencia, pero la mayoría de ellos pareciera ser que no explican el fenómeno en su totalidad, es decir, cuando intentan explicarla tienden a dejar algo afuera. Esto se produce cuando tratamos de lidiar con aquello que Chalmers denomina el 'Problema Difícil de la Conciencia' (1996, p. 26), que corresponde a cómo explicar un estado consciente en relación con su base neurobiológica. Es decir, si el estado neuronal N es la base para una sensación S (por ejemplo, que el estado de activación neuronal en la región occipital del cerebro, en la zona V1, genere la sensación de azul), ¿Qué hace que esa base neuronal en particular, y no otra, genere esa sensación N, o que, de hecho, genere algo en lo absoluto? Lo anterior contrasta con lo que el mismo autor denomina como el 'Problema Fácil de la Conciencia', que correspondería a determinar cuál es la función de la conciencia en el sistema cognitivo.

Cómo enfrentar las dificultades antes mencionadas variará según el grado de importancia que se le entregue al rol de la conciencia dentro de la cognición general. Un tipo de aproximación sostiene que la conciencia, al igual que el resto de los fenómenos en el mundo, es física, por lo que debe ser reducida a la explicación de lo que ocurre en el cerebro (aproximación también denominada como 'Reducciones de Identidad Tipo'), mientras que otras aproximaciones han intentado dar una explicación estableciendo que el problema lisa y llanamente no se puede aprehender, y que en realidad nunca se podrá dar una respuesta satisfactoria, dado los avances científicos actuales, o en su defecto, que en el actual estado del arte de la disciplina no se encuentran las herramientas adecuadas para tales efectos (aproximación denominada 'Misterianismo'); otra manera de resolver el conflicto ha sido obviar el problema estableciendo que la conciencia existe sólo como algo aparente, una propiedad emergente y explicativamente irrelevante para el resto de los fenómenos cognitivos (aproximación denominada 'Eliminativismo'). Finalmente, existen aproximaciones que sí aceptan la realidad psicológica del problema (tales como el 'Realismo Fenoménico' o 'Inflacionismo') y plantean la "necesidad" de dar cuenta del fenómeno toda vez que ésta es un hecho concreto del mundo, y que por lo tanto no puede ser reducido a otro tipo de términos. Para este último tipo de aproximaciones, la conciencia posee un rol causal dentro de la actividad mental, por lo que es menester de la Ciencia Cognitiva entregar una solución empírica del fenómeno.

La posición adoptada en este trabajo corresponde a esta última. Una visión inflacionista del problema sería la más acertada, ya que primero, da cuenta de la realidad del fenómeno

en cuestión sin buscar eliminarlo y reducirlo a otra cosa (ya sea esta reducción empírica o conceptual), y segundo, intenta buscar una explicación concreta a la experiencia subjetiva, es decir, se acerca a la misma como un fenómeno a ser tratado con el mismo estatus que otras características de la actividad mental. Eso ha llevado a algunos autores como Flanagan (1997) a sostener la necesidad de explicar científicamente la conciencia, a través de un programa tanto interdisciplinario como multidisciplinario que permita dar cuenta de la globalidad del fenómeno. Según él, aproximaciones como la reduccionista dejan algo de lado sobre la conciencia al intentar buscar sólo su correlato biológico. No basta saber dónde se produce, ni qué tipo de actividad en el sistema nervioso la genera, ya que una explicación así deja la fenomenalidad de la experiencia relegada a un segundo plano. He ahí la necesidad de una aproximación que dé cuenta, tanto de los aspectos objetivos de la conciencia así como de los subjetivos.

En resumen, sabemos que la conciencia se origina en el cerebro. ¿Cómo le encontramos sentido al hecho de que el disparo de neuronas crea sensaciones, o que produzca emociones adjuntas a determinados estados mentales? Sabemos que el Materialismo dice que todo en el universo es físico y que lo mental consiste en relaciones causales entre estados mentales, *inputs* y *outputs*. El problema es que las teorías físicas son más apropiadas para responder preguntas tales como por qué los sistemas tienen una determinada estructura física y cómo desempeñan ciertas funciones. Pero la pregunta sobre la conciencia es una pregunta distinta, ya que requiere ir más allá de la explicación sobre su estructura y función. No todo el conocimiento de la mente se agota al consignar su descripción física, ni al encontrar todas las relaciones causales y algoritmos de la conducta inteligente, ya que existen ciertos aspectos de la cognición humana que no se explican con una reducción al sustrato físico que las origina. La conciencia posee un aspecto cualitativo, y dada la naturaleza de las teorías físicas que buscan describir funciones y correlatos, la explicación sobre cómo, qué, y por qué se produce la conciencia, quedan inconclusas.

En segundo lugar, existe otro problema, sobre el cual pondremos especial atención en las siguientes secciones; aquél de la confusión sobre qué es la conciencia. Esto último ha generado problemas en relación con su modelamiento así como en cuanto a su estudio. Esa ha sido la razón para que muchos investigadores caigan en algún tipo de confusión, ya que se enfocan en distintos significados de la palabra; por ejemplo, en I.A., se plantea que hace varios años ya pudieron entregar conciencia a una máquina, siendo conciencia entendida en este caso como auto-regulación y *self-monitoring*. Se hace preciso señalar y aclarar filosóficamente con mayor acuciosidad a qué referimos cuando hablamos de conciencia.

3.2 Hablamos de Esto: Dos Aproximaciones a la Conciencia

Antes de continuar, se hace necesario precisar con claridad cuál es el concepto de conciencia al cual referimos en este trabajo. Este será definido por los planteamientos de dos autores, Thomas Nagel (1937 -) y Ned Block (1942 -), con sus conceptos '*What is it like to be'* y la distinción 'Conciencia de Acceso v/s Conciencia Fenoménica', respectivamente.

3.2.1 Nagel y '¿Qué se Siente ser un Murciélago?'

En el año 1974, Nagel escribió el celebre artículo llamado '*What is it like to be a bat?*'. El contexto en el cual fue escrito fue durante el auge de la posición reduccionista, la cual sostenía que si lo único que existe es lo material, la actividad mental debe ser explicada en términos físicos. Pero es un hecho que los humanos, y probablemente otros animales, tengan experiencias conscientes, lo que significa que existe algo así como la "experiencia subjetiva de ser" ese organismo. Es decir, existe una manera de ser, sentir y experimentar de acuerdo con cada ser vivo que tiene una estructuración física lo suficientemente compleja como para desarrollar una manera de sentirse esa criatura en particular. El autor refiere a ésta última, para luego comentar el error en el cual una aproximación reduccionista incurriría:

Podríamos llamar a éste último el carácter subjetivo de la experiencia. Éste no se pude capturar por ninguno de los análisis que recientemente se han articulado para reducir la mente, ya que todos ellos son compatibles lógicamente con su ausencia. [El carácter subjetivo de la experiencia] no es analizable en términos de ningún tipo de explicación de estados funcionales o de estados intencionales, [...] y no es analizable en términos del rol causal de las experiencias en relación con la conducta humana típica, por razones similares. (Nagel, 1974, p. 519, traducción mía.)

En otras palabras, Nagel sostiene que el problema de un análisis reduccionista sobre la experiencia subjetiva de cada individuo no agota el concepto, ya que deja "algo" fuera, y ese algo es la experiencia subjetiva. En ningún caso se niega, sin embargo, el hecho de que los estados mentales conscientes sí pueden relacionarse causalmente con la conducta, así como también pueden ser analizados funcionalmente. Es decir, la objetividad de un análisis reduccionista o funcionalista de "lo que se siente ser" una criatura es posible, incluso pertinente, pero éste por sí sólo no puede dar cuenta de la otra dimensión de esa experiencia, esto es, la experiencia que involucra un punto de vista subjetivo, ya que un análisis objetivo externo deja fuera aquello que Damasio (2000) etiqueta como la "Presencia de un Alguien", justamente de alguien que *experimenta* el mundo.

Es esta discusión la que lleva a Nagel a entregar el ejemplo de un murciélago. Existe una experiencia de ser murciélago, y ninguna apreciación desde la tercera persona podrá saber qué se siente ser un murciélago. Se puede, dice el autor, imitar la conducta de uno, e implementar los dispositivos de recepción de *input* de este mamífero, pero eso no es suficiente para sentir qué se siente ser un murciélago, ya que con ello solamente se entendería qué se siente para un humano ser un murciélago, pero nunca qué se siente para un murciélago ser un murciélago. La razón radica en el hecho de que cada especie se encuentra limitada por su propia mente y experiencia subjetiva, lo que impediría el éxito de tal ejercicio.

No obstante, algún filósofo funcionalista dirá que siempre sería posible establecer experiencias *tipo (type)* para un murciélago basándose en su conducta, su estructura y dinámica de funcionamiento. Según este filósofo, sería posible establecer nociones de frío, dolor y tendencias conductuales, dado un determinado contexto. Lo anterior es cierto, pero nuevamente, siempre existirá un aspecto subjetivo que no es aprehendido ni es posible concebir. Lo anterior se debe a que para concebirlo y aprehenderlo nos encontramos limitados por nuestra propia experiencia subjetiva, lo que hace imposible para un humano saber qué se siente ser un murciélago (Más sobre este punto en el capítulo 4).

Como antes se mencionara, el texto original de Nagel fue escrito en el auge de las teorías reduccionistas de la mente. No obstante, es posible leer entre líneas la presentación

de un obstáculo para el Funcionalismo también. Y tiene que ver con lo siguiente: Dada la distinta naturaleza de las maneras subjetivas de ser criaturas (es decir, por encontrarnos sujetos y confinados a nuestra manera de experimentar el mundo), es posible que un intento funcionalista de explicar la experiencia subjetiva no sea capaz de conceptualizar ciertas cosas que están presentes en ésta. Por ejemplo, el águila, dada su fisonomía, posee un espectro cromático superior al humano, además de exhibir una capacidad de distinguir pequeños objetos a grandes distancias[30]. Por lo mismo, el águila posee ciertas experiencias subjetivas relacionadas con la visión las cuales nos sería imposible siquiera concebir, debido a que nos encontramos limitados por nuestra propia experiencia. En otras palabras, el comentario se vincula con la necesidad de dejar de lado un cierto "chauvinismo humano" con respecto a la sistematización de las experiencias subjetivas presentes en otras especies - y por extensión, en criaturas artificiales - por parte del Funcionalismo, ya que eso puede llevar a olvidar el hecho de que hay ciertos aspectos fenoménicos que jamás será posible representar ni "funcionalizar" con respecto a otras especies o autómata.

En definitiva, el autor llama a las ciencias a considerar el hecho de que existe una experiencia cualitativa consciente acerca de qué se siente ser una criatura en particular. Cualquier intento de reducir esta característica mental a su correlato biológico involucra caer en un error, ya que al momento de volverla objetiva y accesible desde el exterior, se aleja de su objeto de estudio, su carácter intrínseco y subjetivo. En otras palabras, y como Flanagan (1997) señalara décadas después, un estudio de la conciencia no debiera negarse al hecho de que la conciencia se siente de determinada manera, y que a diferencia de otros aspectos cognitivos en los cuales se busca desmantelar las apariencias del fenómeno para desnudar su realidad, la conciencia debe ser "el estudio sobre las apariencias", porque en el caso de la experiencia subjetiva, la apariencia *es* la realidad. Esa realidad está directamente asociada a una manera de ser de cada ser vivo, y por lo tanto, no se debe negar el hecho de que existe una manera de ser ese ser vivo al momento de dar cuenta de su actividad mental. El Funcionalismo, dada su naturaleza objetiva y principios básicos antes mencionados, es incapaz de dar cuenta de esta experiencia subjetiva que cada ser vivo posee, ni tampoco explica cómo ni por qué surge.

3.2.2 Conciencia de Acceso y Conciencia Fenoménica.

Una aproximación similar a la de Nagel es la que toma Block (1995b) en su artículo sobre lo qué él considera una confusión sobre la conciencia. El problema, según el autor, radicaría en el hecho de que la conciencia sería un concepto "mestizo", es decir, que la palabra connota una variedad de conceptos y denota un número de distintos fenómenos, lo que a su vez generaría una serie de malentendidos tanto inter- como intra-disciplinariamente, toda vez que una variedad de conceptos serían tratados como si fueran uno. Lo anterior llevaría a que se produzca algo similar a la 'Fábula de los 6 Hombres Ciegos y el Elefante', en la cual cada uno de los hombres se encuentra en contacto con una parte del elefante (por ejemplo, a uno le toca la pata trasera izquierda, a otro la trompa, a otro el colmillo de marfil, etc.) y cada uno de ellos cree que está tocando la parte esencial del mamífero en cuestión. Es definitiva, según el autor la palabra 'Conciencia' refiere a muchos conceptos diferentes y fenómenos que han sido amarrados como si fueran uno, y eso resulta problemático al

[30] La mayoría de las aves rapaces poseen fóveas con más bastones y conos que la fóvea humana, aproximadamente 65.000 y 38.000 mm^2, respectivamente. La fóvea misma puede ser de forma de lente, incrementando adicionalmente la densidad efectiva de receptores. Esta combinación de factores les otorga una visión a distancia 6 a 8 veces más potente que la de los humanos, permitiéndole ver detalles que nosotros jamás podríamos.

analizar ciertos aspectos de la conciencia usando premisas que se aplican a otros aspectos de ésta.

Lo anterior lleva al autor a dividir la conciencia en distintos conceptos, particularmente, lo que él denomina 'Conciencia Fenoménica' (en inglés abreviada como *P-consciousness*) y 'Conciencia de Acceso' (en inglés *A-consciousness*). Estos conceptos constituyen una distinción básica y necesaria para evitar confusiones. La distinción surge de la idea central de que las propiedades fenoménicas son distintas de las propiedades funcionales de la conciencia. Las propiedades fenoménicas son propiedades de la experiencia subjetiva (o 'propiedades *P-conscious'*). Habría 'estados *P-conscious'*, los cuales incluyen los estados subjetivos de nuestras experiencias sensoriales (visión, audición, etc.) dolor, placer, etc. Es a este tipo de conciencia al que el presente trabajo se dirige al momento de analizar la plausibilidad de que una máquina exhiba conciencia fenoménica.

Por otro lado, encontramos el concepto de *A-consciousness* o conciencia de acceso. Este tipo de concepto corresponde a la concepción no-fenoménica de la conciencia, y se encuentra relacionada con las tareas cognitivas de representación y control de la conducta. Un estado mental es *A-conscious* si éste se pone a disposición del aparato cognitivo para el funcionamiento inteligente y para el control de la acción; es decir, para que un estado sea *A-conscious*, éste debiera encontrarse disponible para su reportabilidad (o capacidad de ser considerado como un elemento en el funcionamiento mental) en caso de que el sujeto lo disponga.

Block considera que tanto *P-consciousness* como *A-consciousness* interactúan entre sí, y sugiere la posibilidad de que ocurran simultáneamente. Sin embargo, también es posible encontrar casos en las que ocurre una o la otra, dato que el autor utiliza para reafirmar sus postulados con respecto a la dicotomía, y señala la existencia de casos clínicos en los cuales existe una sin la otra. (pp. 385-87, Op. Cit.).

Es importante detenernos un momento para profundizar algo más con respecto a la distinción *A/P-consciousness*. Con respecto a *P-consciousness*, es necesario señalar que ésta se asemeja a lo que Nagel denominara *'What It Is Like to Be'* o cómo se siente ser determinada criatura. Dada la naturaleza privada de *P-consciousness*, sería imposible generar una definición de ésta, por lo que se remite al hecho de sólo poder "apuntar" a lo que referimos cuando hablamos de *P-consciousness*. Las propiedades *P-conscious* son propiedades relacionadas con la experiencia, y los estados *P-conscious* son estados sobre la experiencia personal del mundo, es decir, un estado es *P-conscious* si es que posee propiedades *P-conscious*, y realizan lo que se denomina "qué se siente tener un determinado estado mental" (también entendida como la experiencia subjetiva de tener ese estado, en la misma línea del concepto de *Qualia* antes mencionado en la sección 2.3.2). Las propiedades *P-conscious* incluyen las propiedades subjetivas de sentir, percibir y captar, además de incluir pensamientos, emociones y otros estados mentales. Sería este tipo de conciencia la que una aproximación reduccionista o eliminativista dejaría de lado. Desde un punto de vista funcionalista, *A-consciousness* y *P-consciousness* serían idénticos, toda vez que esta aproximación no contempla ni explica el hecho que existe una experiencia subjetiva del mundo, además de considerar que la mente puede ser completamente modelada sólo en función de las relaciones causales entre sus componentes. Es decir, se produce una reducción del aspecto fenoménico de la conciencia sobre las funciones que ésta tiene, dejando un aspecto sin ser considerado.

Con respecto a la conciencia de acceso o *A-consciousness*, se señala que ésta es la que produce la mayor cantidad de confusiones al tratar indistintamente la dimensión fenoménica (*P-consciousness*) con la dimensión no-fenoménica de la conciencia (*A-*

consciousness). Un estado es *A-conscious* si es que éste es transmitido al resto del sistema cognitivo y está disponible para la acción cognitiva. Un estado *A-conscious* tiene una representación de acceso que es la que le permite actuar causalmente con el resto del sistema cognitivo. Por tanto, *A-consciousness* juega un rol en nuestra vida diaria, y podría ser considerada como el correlato de procesamiento de información de *P-consciousness.*

A manera de diferenciarlas claramente, el autor sostiene que existirían 3 diferencias entre ambos conceptos de conciencia. Primero, el contenido *P-conscious* es fenoménico, mientras que el contenido *A-conscious* es representacional, es decir, el sistema lo transforma de manera tal que pueda ser utilizado por la mente en el desempeño cognitivo. Sin embargo, según el autor es importante recordar que:

El contenido de una experiencia puede ser tanto P-conscious como A-conscious, el primero en virtud de sus experiencias fenoménicas y el segundo en virtud de sus propiedades representacionales. (p. 383, Op. Cit., traducción mía.)

Es decir, el autor sostiene que cuando referimos a la conciencia, ésta posee una bi-dimensionalidad determinada por sus propiedades que se ve reflejada en los dos tipos de conciencia. En este sentido, *A-consciousness* y su rol en el razonamiento lleva a la segunda diferencia entre ambas: *A-consciousness* es una noción funcional, es decir, se involucra en el funcionamiento de un sistema ejecutivo central para el control de la acción, y lo que hace que un estado sea *A-conscious* es lo que su contenido hace en el sistema, mientras que *P-consciousness* no es una noción funcional. Y tercero, existen estados *P-conscious* que son *tipo (type)* o clases de estados, por ejemplo, la sensación de dolor es un *tipo P-conscious*, ya que cada dolor debe sentirse de esa manera. Mientras que estados *A-conscious*, son *tokens* orealizaciones, es decir, se encuentran accesibles para la acción inteligente en determinado momento, y no pueden ser usados recursivamente. En resumen:

El paradigma de estados P-conscious corresponde a las sensaciones, mientras que el paradigma de estados A-conscious corresponde a estados de 'Actitudes Proposicionales' tales como pensamientos, creencias y deseos, es decir, estados con contenido representacional expresado en cláusulas 'que+oración'. Sin embargo, como dije antes, los pensamientos son usualmente P-conscious y las experiencias perceptuales comúnmente tienen contenido representacional. (p. 384, Op. Cit., traducción mía.)

Con respecto a la interacción entre ambos conceptos de la conciencia, es posible proponer que éstas influyen una sobre la otra, toda vez que los estados *P-conscious* pueden servir como base para la generación de estados *A-conscious*, así como estados *A-conscious* pueden permitir una apreciación más fina del contenido fenoménico (p. 400).

Por lo tanto, es posible evaluar la distinción de Block, estableciendo que su principal característica es que permite explicar el problema de la conciencia, buscando una aproximación cognitiva para la conducta sin excluir propiedades fenoménicas en esas explicaciones. Ahora bien, el autor admite que la relación entre éstas no es del todo clara, pero que independientemente de generar una teoría acabada de la conciencia, el objetivo es diferenciar ambos aspectos que normalmente son confundidos. Y es precisamente sobre esta confusión entre el aspecto fenoménico y funcional de la conciencia donde la I.A. encuentra sus mayores problemas. Es por eso que es pertinente referir en la siguiente sección a dos áreas esenciales para el trabajo de *'Machine Consciousness'* (Alexander, 2007): Por qué la I.A. debiera interesarse en modelar máquinas fenoménicamente conscientes, para luego dar cuenta del trabajo reciente en los últimos años en el campo de *Machine Consiousness*.

3.3 Conciencia + I.A. = Explorando el Concepto de '*Machine Consciousness*'.

Como ya se ha mencionado, en los últimos años ha habido un resurgimiento en el interés por el estudio de la conciencia desde diversas disciplinas tales como la Psicología, la Filosofía y la Neurociencia. Los científicos buscan entender la conciencia como un producto de la máquina más compleja de todas: El cerebro. Dicho interés también se ha ramificado a la I.A. y puede ser resumido en la presentación de tres propósitos: Primero, ser capaces de diseñar modelo(s) de la conciencia; segundo, que las implementaciones de esos modelos puedan ser útiles para comprender qué es; y tercero, implementar un modelo de máquina consciente que permita aportar al conocimiento científico de la misma. La relación máquina-cognición posee una larga data, y eso se refleja en lo que Newell (1980) define como el rol básico de un sistema físico de símbolos para el estudio de la mente humana:

La hipótesis es que los humanos son instancias de sistemas físicos de símbolos, y, a través de esto, las mentes entrar en el universo físico [...] Esta hipótesis fija los términos en los cuales buscamos una teoría científica de la mente. (p. 136, traducción mía.)

Por otro lado, la premisa detrás de la incorporación de una "conciencia" en determinadas máquinas se debe a la posibilidad de que ésta podría llevar a crear máquinas más inteligentes. A este tipo de investigación se le ha denominado '*Machine Consciousness*' o conciencia de máquina (Gámez, 2007), también conocido como '*Machine Modeling of Consciousness - MMC*' o modelamiento de la conciencia utilizando máquinas (Alexander, 2007). El objetivo de los investigadores relacionados con MMC podría resumirse en un objetivo doble: Descubrir la naturaleza de la conciencia fenoménica (relacionado con lo que Chalmers denominara 'El Problema Difícil de la Conciencia') y el rol de la conciencia en el control de la conducta, además de comprender cómo ésta se relaciona con otras capacidades cognitivas como la atención, la planificación, etc.

Ambos aspectos del trabajo en MMC se encuentran conectados, y debido a eso existen diversos criterios o programas de investigación desde los cuales diseñar máquinas conscientes. Por ejemplo, algunos programas se enfocan en replicar la conducta asociada a la conciencia, o bien otros investigadores se interesan en entregar propiedades fenoménicas a las máquinas. Es por eso que Gámez (2007) considera útil delimitar el trabajo en *Machine Consciousness* de acuerdo con los siguientes criterios:

- MC1: Máquinas que exhiben la conducta externa asociada a la conciencia.

- MC2: Máquinas con las características cognitivas asociadas a la conciencia.

- MC3: Máquinas con una arquitectura que asegura ser la causa o correlato de la conciencia humana.

- MC4: Máquinas con conciencia fenoménica.

Por supuesto, es posible que ciertos programas consideren más de un criterio. Por ejemplo, un programa que busque MC4 lo puede hacer a través de arquitecturas basadas en funcionamiento neuronal, es decir, MC3. Una clasificación de esta naturaleza permite delimitar el área de investigación hacia la cual se dirigen las críticas cuando se sostiene que una máquina programada no puede exhibir conciencia fenoménica. Para propósitos de este trabajo, las críticas se encontrarían en directa relación con el trabajo en MC4, además de permitir distinciones claras en cuanto a que replicar MC1 no es replicar MC4.

Se hace pertinente ahondar un poco más con respecto a las características de cada uno de los criterios utilizados en *Machine Consciousness*. Primeramente, el trabajo relacionado con MC1 realiza la distinción entre las conductas observables que son conscientes y aquellas que no requieren de conciencia (tales como las contracciones musculares al momento de saltar), o bien actividades que han logrado ser automatizadas (tales como atar los cordones de nuestros zapatos). Sin embargo, una gran parte de nuestra conducta *sí* se encuentra asociada a nuestra conciencia, especialmente tareas complejas que requieren nuestra atención y se nutren de nuestra percepción (tales como tomar apuntes durante una clase). Una distinción entre estos dos tipos de conducta ayuda a establecer que MC1 busca implementar este último tipo de actividad. Es importante aclarar, no obstante, que si bien las conductas a ser replicadas podrían requerir de conciencia fenoménica en la mente humana, esto no es motivo de interés en MC1, ya que siempre será posible contemplar el hecho de que éstos sean robots *zombies,* i.e., máquinasque llevan a cabo tareas sin experimentar absolutamente nada desde un punto de vista fenomenológico. Por lo mismo, el criterio para adscribir dicha conciencia se encuentra limitado solamente a la conducta observable de la máquina. Es decir, el Test de Turing para MC1 es considerado como condición suficiente para determinar si una máquina posee o no conciencia.

Se denomina MC2 - máquinas con las características cognitivas asociadas a la conciencia - al área que busca conexiones entre la conciencia y fenómenos cognitivos tales como la imaginación, la creatividad, el aprendizaje y otros. Es este tipo de investigación la que menos se relaciona con la implementación de conciencia fenoménica (MC4), toda vez que es posible simular miedo o creatividad en una máquina sin que ésta requiera la presencia fenomenológica de la misma. En algunos casos, el modelamiento de facultades cognitivas apunta a un desempeño lo más realista posible (MC1) o utilizando arquitecturas asociadas a la conciencia (MC3), pero también es posible concebirlo sin interactuar con ninguna de estas áreas, por ejemplo, en un programa que no exhiba conducta observable basado en algoritmos simples. Existe un área relacionada a MC2 que goza de gran popularidad por estos días y corresponde a lo que se ha denominado ‘Sistemas Expertos’, entendidos como una aplicación informática capaz de solucionar un conjunto de problemas que exigen un conocimiento específico y aplicado a determinados temas. Un sistema experto es un conjunto de programas que, sobre una base de conocimientos, posee información de uno o más expertos en un área específica. Lo anterior es un ejemplo de que es posible replicar una capacidad (en este caso, generar predicciones o entregar soluciones a determinados problemas) sin necesidad de recurrir a ninguna de las otras 3 áreas.

MC3 refiere al trabajo realizado en función de las arquitecturas responsables de producir la actividad consciente, y surge como un intento de modelar y probar teorías sobre ésta. Entre estas teorías, se encuentra el trabajo en función de una arquitectura denominada ‘Global Workspace’ (Baars, 1993) que permite la disponibilidad y transmisión de información a un espacio al cual la conciencia y otros aspectos de la actividad mental pueden acceder. MC3 se traslapa tanto con MC1 como MC2 si es que se busca reproducir características conductuales o cognitivas, pero también podría relacionarse con MC4 si es que se considera que una determinada arquitectura puede producir estados fenoménicos. Ahora bien, siempre es posible contemplar el hecho de que simular una arquitectura “consciente” en una máquina no signifique que la máquina sea, de hecho, consciente (Más sobre esto en Capítulo 4).

Finalmente, MC4, que resulta ser el área más controversial. Esto se debe a que MC1, MC2 y MC3 se encuentran de una u otra manera relacionadas con lo que Block denomina *A-consciousness*, por lo que no declaran en ningún caso crear u otorgar estados fenoménicos

a una máquina. MC4, por otro lado, presenta problemas filosóficos concretos, toda vez que consiste en entregar experiencias fenoménicas reales a una máquina. Apelando a la distinción entre 'Inteligencia Artificial Fuerte' e 'Inteligencia Artificial Débil'[31], sería posible asociar el trabajo en MC4 con la primera, mientras que el trabajo en MC1, MC2 y MC3 con la segunda. Es decir, el trabajo que se lleva a cabo en MC1-MC3 se relaciona directamente con una aproximación objetiva y funcional de la conciencia, por lo que se enfocan en el estudio de *A-consciousness*, mientras que MC4 se relaciona directamente con *P-consciousness*, ya que es la única de las áreas que afirma estar tratando directamente con la experiencia cualitativa en máquinas.

La relación de MC4 con las otras áreas de investigación en más bien difusa, debido a que puede encontrarse relacionada estrechamente con una arquitectura determinada, o con un fenómeno cognitivo específico o con conductas asociadas a la conciencia. Al mismo tiempo, puede desarrollarse en distintos sistemas inspirados en otras concepciones de la mente, como por ejemplo, desde los planteamientos de la Cognición Corporalizada[32], la Cognición Situada[33] u otras, lo que refleja el hecho de que el trabajo en MC4 requiere de compromisos teóricos fuertes por parte de investigadores que aseguran trabajar en la creación de máquinas que, efectivamente, sienten.

A pesar de las dificultades que MC4 posee, existe una línea de trabajo en la actualidad que trata de dar cuenta de las experiencias fenoménicas en máquinas sin buscar una reducción de *A-consciousness* sobre *P-consciousness.* En esa línea, el trabajo de Alexander (2007) plantea una perspectiva en la cual intencionalidad y conciencia fenoménica se encuentran vinculadas. Particularmente, su propuesta busca formalizar la brecha entre la experiencia visual desde la primera y la tercera persona en el contexto de la I.A. Su aproximación se denomina 'Teoría Axiomática de la Conciencia'. Ese trabajo es uno de los pocos que se enfocan explícitamente en la conciencia fenoménica a través de una aproximación híbrida que combina análisis introspectivo de la experiencia subjetiva y modelos computacionales. Su modelo posee la ventaja de poder ser considerada como un punto de partida para el trabajo en MC4, dado que la mayoría del trabajo en *Machine Consciousness* se enfoca en MC1-MC3.

Ahora bien, y adentrándonos en el ámbito de la conciencia fenoménica en máquinas, cabe preguntarse qué hace que una máquina tenga conciencia. Las respuestas variarán dependiendo de la arquitectura usada, ya sea la Arquitectura Clásica o la Conexionista[34]. Desde el paradigma clásico, una máquina exhibirá cualquier fenómeno cognitivo si es que los estados mentales de ésta son modelos simbólicos del mundo, en los cuales la computación que lleve a cabo esté explícitamente representada en sus algoritmos. Desde el paradigma conexionista, lo relevante son los mecanismos que son responsables de tales representaciones a través de modelos no-simbólicos. No obstante, determinar qué hace a

[31] Distinción planteada por Searle (1980) referida en la sección 2.3.

[32] Los investigadores de distintas áreas que adscriben a una concepción corporalizada de la mente sostienen que su naturaleza y la manera cómo opera se encuentra determinada en gran medida por la estructura del cuerpo humano. Estos aspectos incluyen el sistema de percepción, el conocimiento que los seres humanos desarrollan con el tiempo con respecto a la 'propiocepción' - o capacidad de poder ubicar el cuerpo en el espacio, también conocido como la manera que tenemos de conocer nuestra postura con los ojos cerrados - en relación con la capacidad de movimiento, y las interacciones con nuestro medio ambiente que van moldeando la manera cómo lo enfrentamos.

[33] La Cognición Situada es el estudio de la cognición en su ambiente natural. Esta perspectiva enfatiza que la mente subjetiva opera en medio ambientes que estructuran, dirigen y moldean los procesos cognitivos.

[34] Se desarrolla sobre estas arquitecturas en extenso en la sección 2.2.2.

una máquina tener experiencias fenoménicas permite diferenciar claramente cuándo sería posible determinar si una máquina efectivamente tiene esas experiencias y cuándo no. Dado el actual estado del arte de la disciplina, dicha pregunta está lejos de ser respondida con claridad. Según Alexander (Op. Cit., p. 88), por estos días existen trabajos tanto en la línea clásica como la conexionista que buscan entregar mayor precisión con respecto a esto, de manera tal de discernir desde la tercera persona la presencia de estados fenoménicos en otros seres vivos o máquinas. Sin embargo, no existen parámetros claros o condiciones necesarias y/o suficientes en la actualidad que permitan aseverar la presencia/ausencia de conciencia en una máquina, por lo que mayor investigación es necesaria.

Existe una pregunta esbozada al principio de este capítulo que si bien sencilla, es básica y necesaria: ¿Por qué la I.A. Fuerte debiera interesarse en entregar conciencia fenoménica a una máquina?

Existen dos razones para ello. Primero, el trabajo en I.A. serviría para clarificar la noción de "qué se siente estar consciente" y cómo eso mejoraría el desempeño de una máquina en tareas inteligentes. Según la disciplina, cualquier cosa que pueda ser sintetizada[35] puede ser construida, y si el resultado de esta construcción logra entregar conciencia (o algún grado de ésta) a una máquina, entonces sería sensato establecer que existe la posibilidad de que sea consciente, y muy probablemente, la presencia de algún grado de conciencia se verá reflejado en cambios o mejoras de su conducta, lo que eventualmente llevará a crear máquinas que lograrán un mejor desempeño que otra que no la tenga. En esta línea de investigación, la actual generación de robots diseñados para interactuar con humanos ha mostrado un buen resultado con respecto a su funcionamiento mecánico y el control de sus movimientos, pero éstos presentan "capacidades limitadas de percepción y razonamiento, además de poca capacidad de acción en ambientes poco estructurados o cambiantes" (Chella y Manzotti, 2007). Una nueva generación de robots que busque, por ejemplo, interactuar con humanos en medios ambientes cambiantes y normales, necesitarán una mejor capacidad para "darse cuenta" del ambiente a su alrededor, de sus eventos y objetos, es definitiva, de un medio que se modifica a cada instante. Es decir, una nueva generación de robots necesitaría una forma de conciencia artificial que le permita dar cuenta de esos cambios. En esa misma línea argumentativa, Alexander (Op. Cit.) sugiere que las mejoras que una máquina consciente exhibiría se verán reflejadas en las siguientes características:

[...] una mejor autonomía, libertad con respecto a su pre-programación, y habilidad para representar su propio rol en el ambiente. Esto mejoraría la capacidad para la acción basada en una actividad 'contemplativa' interna más que en una acción reactiva basada principalmente en una tabla que contenga acciones de contingencia pre-programadas. (p. 89, traducción mía.)

En otras palabras, una máquina consciente, capaz de sentir y de reflexionar sobre el mundo, posee ciertas características que le permitirían ir más allá del algoritmo para desarrollar determinada tarea. La conducta inteligente podría verse asociada no sólo con el correcto funcionamiento de un grupo de reglas establecidas y bien formuladas, sino también con la interacción de esta "capacidad cognitiva de ser consciente" que le facultaría - entre otras cosas - alterar sus programas en favor de una mejor respuesta al ambiente, determinando la importancia del rol de la conciencia para desarrollar ciertas tareas.

Es sabido que la mayoría de los mamíferos, particularmente los humanos, parecen exhibir tanto *A-consciousness* como *P-consciousness*. Por lo tanto, es razonable suponer

[35] Definida como el acto de componer o crear un cuerpo a partir de sus elementos separados en un proceso de análisis previo.

que las arquitecturas cognitivas responsable de éstas posean alguna ventaja evolutiva. De lo anterior se desprende que la conciencia juega un rol en la cognición humana, y que ésta no es sólo un epifenómeno o un elemento inoperante sin un rol causal en el sistema. Piense, por ejemplo, en el rol que la experiencia subjetiva tendría en tareas que requieran empatía por parte de un interlocutor, o tareas en las cuales la percepción sensorial posea un papel preponderante (más sobre esto último en capítulo 4).

Lo que nos lleva a la razón número dos del por qué la I.A debiera crear máquinas conscientes: La actividad inteligente humana *sí* requiere conciencia para una gran cantidad de tareas que requieren un funcionamiento cognitivo, tales como la atención, la participación en la creación de un *self*, la acción guiada, la autonomía, el aprendizaje, las emociones, la amplia gama de experiencias fenoménicas que nos permiten sentir y experimentar el mundo, que varían desde el placer de la música hasta la ventaja evolutiva de evitar el dolor en determinadas situaciones (Damasio, 2000). Autores como Baars (en Block, 1995b, p. 401) sugieren alrededor de 18 funciones tanto para *A-consciousness* como *P-consciousness*. El ejemplo más común proviene de casos en los cuales se produce uno de los fenómenos más interesantes y novedosos en el área, aquél denominado *Blindsight* [36]. El fenómeno refleja que la función de la conciencia sería permitir que la información de los sentidos pueda ser usada para algún tipo de acción controlada, y que la presencia/ausencia de ésta muestra claras diferencias conductuales. Por ejemplo, si un paciente *blindsight* que presenta una adicción al cigarrillo quiere desesperadamente fumar, y el cigarrillo se encuentra dentro del campo visual dañado, el sujeto no irá a buscarlo, no importando cuánto desee fumar o si percibe la presencia del cigarrillo a través de otro sentido como el olfato. Por lo tanto, es posible concluir que la experiencia consciente facilita la comprensión y nuestro buen desempeño en el medio ambiente que nos encontramos, y dado que este tipo de pacientes no posee la experiencia subjetiva de ver el área dañada, esto se traduce en un tipo de conducta que lleva a nuestro paciente a no saciar su adicción. En otras palabras, la información debe "estar" dentro de nuestra *A-consciousness* para que desempeñe el rol de guiar nuestras acciones voluntarias, considerando el hecho que para que eso ocurra, primero debe haber una *P-consciousness* que recoja esa información. De la misma manera, la sensación subjetiva de ansiedad que un fumador compulsivo privado de tabaco también juega roles significativos en cuanto al tipo de conducta que este podría tener en determinadas situaciones, ya que existe una propiedad cualitativa asociada a la ansiedad de no poder fumar.

Por otro lado, la conciencia resulta ser esencial para el correcto desempeño de los individuos en una de las tareas más difíciles y complejas que caracteriza a nuestra especie: El desenvolverse socialmente entre sus pares y navegar la dimensión social y cultural en la que nos desenvolvemos, lo que también se denomina 'Cognición Social' (Robbins, 2008). Según el autor, la conciencia fenoménica y la dimensión social a la que nos encontramos sujetos son capacidades que se encuentran interconectadas, toda vez que nos permiten navegar mejor el mundo. Considere, por ejemplo, cómo permite entender los estados mentales de otro sujeto que dice encontrarse "sufriendo" o padeciendo algún tipo de malestar, toda vez dicho estado posee un componente cualitativo al cual sólo

[36] *Blindsight* se define al fenómeno en el cual personas que exhiben ceguera en alguna zona de su campo visual pueden, no obstante, presentar respuesta a ciertos estímulos visuales en el área dañada. Por ejemplo, un paciente con dicha condición no pueden darse cuenta o percibir (no hay *P-consciousness*) ningún estimulo en el campo visual dañado, sin embargo, pueden predecir con un alto grado de probabilidad si es que se les da a elegir entre dos opciones con respecto al estímulo, por ejemplo, en términos de dirección, ubicación o tipo de movimiento. Lo anterior, según autores como Block (1995b), serviría para hacer manifiesta la distinción *A/P-consciousness*.

podemos acceder desde la experiencia subjetiva cualitativa personal de encontrarse en un estado similar. Es decir, puedo comunicarme y generar empatía con los sujetos a mi alrededor a través de la experiencia cualitativa personal de encontrarme en esos estados. Además, sería posible entender fenómenos tales como el 'Dolor Social' (Robbins, 2008, p 16), que en términos generales, corresponde a los estados asociados a la percepción del posible daño a las relaciones interpersonales. Lo anterior se conectaría con lo que se denomina 'Contagio Afectivo' (Robbins, Op. Cit, p. 17), que representa la tendencia que poseen las emociones, estados de ánimo y afectivos a ser propagados entre personas en contextos sociales determinados. De acuerdo al autor, distintas investigaciones sobre estos fenómenos sugieren que la conciencia afectiva depende de la percepción del mundo social de la misma manera que depende de la percepción del cuerpo, sugiriendo la idea de que la conciencia se encuentra socialmente "corporalizada". En otras palabras, una aproximación como la de este autor contempla la dimensión socio-cultural en la cual nuestra mente se desenvuelve, y que el carácter social humano depende de la habilidad para representar los estados conscientes de otras personas. La conciencia fenoménica, por tanto, jugaría un rol crucial para el correcto desempeño en tareas cotidianas como la representación de estados mentales de otras personas en tareas como una conversación o interacción interpersonal.

En conclusión, la I.A. desde sus inicios con el trabajo de Turing ha evitado enfrentar el tema de la conciencia, y es sólo hasta un par de décadas que ha resurgido interés en ésta. Según autores provenientes desde distintas disciplinas, la conciencia se relaciona con muchos aspectos de la cognición humana que son esenciales para la actividad inteligente y otros aspectos tanto cognitivos como sociales. Simultáneamente, la Inteligencia Artificial Clásica enfrenta muchos problemas con respecto a cómo diseñar máquinas conscientes en función de la premisa de que "el algoritmo correcto es todo lo que se necesita para capturar la actividad inteligente". Las dificultades se producen al momento de replicar lo que el sistema nervioso humano realiza: La coordinación entre el funcionamiento del cerebro, el cuerpo y el ambiente en el cual se desenvuelve. Es la conciencia, (tanto *A-consciousness* como *P-consciousness*) la que permite la posibilidad de tener la experiencia subjetiva de ser un individuo al cual le ocurren cosas, lo que al mismo tiempo permite un mejor desempeño en aquello que denominamos actividad inteligente, y que dicho sea de paso, ningún algoritmo proporciona ni explica. Es por eso que es posible generar una serie de consideraciones críticas con respecto a la plausibilidad de que una máquina programada en función de algoritmos pueda poseer conciencia fenoménica.

4. Conclusiones: La Implausibilidad de que la Programación a través de Algoritmos entregue Conciencia Fenoménica a una Máquina.

Piense en la siguiente situación: Ud. se despierta en el medio de la noche, en una habitación silente y oscura en una cabaña en el medio de un bosque, aislado. Ud. está despierto, totalmente consciente, y no recibe ningún tipo de estímulo sensorial relativamente importante (quizás, si se esfuerza, pueda sentir el peso de las sábanas sobre su pecho). Está despierto sin estar necesariamente realizando ningún tipo de tarea, ni tampoco recibe estímulo alguno que lo lleve a una acción voluntaria. ¿Es posible entregar este tipo de conciencia a una máquina programada en función de algoritmos? ¿Es plausible otorgar experiencias subjetivas a un aparato diseñado y construido por el hombre?

La conciencia fenoménica corresponde a la experiencia subjetiva y personal que todo individuo siente al desenvolverse en el mundo. Esta idea de conciencia se ha confundido en la literatura sobre el tema con otras nociones similares, tales como el *self* o lo que Block (1995b) denomina *A-consciousness,* que corresponde a la concepción de conciencia involucrada en los procesos mentales para la acción inteligente, algo así como la *representación mental* de la conciencia fenoménica que se relaciona causalmente con el resto de la mente. Pero la conciencia fenoménica corresponde a lo que se siente para cada especie al experimentar el mundo de determinada manera, lo que Nagel (1974) etiqueta como 'Qué se Siente ser Determinada Criatura'. En una línea argumentativa similar, Chalmers (1996) refiere como *Qualia* a las propiedades cualitativas que acompañan a determinado estado mental. Es este tipo de conciencia la cual resulta problemática para el estudio de la mente, y es esta característica la que resulta difícil de entregar a una máquina programada en función de algoritmos.

La I.A. se ha caracterizado por abrazar una concepción funcionalista de la mente, y apela a un diseño el cual plantea que la generación de actividad inteligente se logra a través de la implementación del algoritmo correcto para determinada tarea. Esta idea se origina desde el concepto de Máquina de Turing, y según el Funcionalismo, ésta define en términos muy generales qué es pensar: Aplicar un algoritmo a un problema y a través de los pasos finitos de éste llegar a un resultado exitoso sin la necesidad de conciencia ni nada externo al mismo algoritmo. Esta visión ha logrado grandes avances en la comprensión de la mente, además de entregar un programa de investigación que, simultáneamente, propició la separación entre mente y su correlato físico, haciendo admisible la posibilidad de crear actividad inteligente en artefactos que no estén compuestos de las mismas estructuras biológicas que los seres vivos. Incluso, sería posible expandir el comentario a los planteamientos de una Arquitectura Conexionista (recordando que esta última exhibe diferencias importantes en comparación con la Arquitectura Clásica - ver capítulo 2), toda vez que ambas corresponden a una concepción representacional-computacional de la mente: Ya sea a través de la metáfora del computador o la metáfora del cerebro, el

estudio de la cognición implica una reducción a un sistema computacional. Aquello, sin duda, trae consecuencias sobre los fenómenos a los que se pretende dar una respuesta, y he ahí la dificultad de la I.A. al tratar de explicar e implementar la conciencia: Una aproximación computacionalista corresponde a una visión desde la tercera persona, es decir, una aproximación objetiva, mientras que la conciencia fenoménica tiene una ontología de primera persona. La experiencia subjetiva de un individuo sólo es accesible por él mismo, y esa característica se pierde si se la pretende estudiar desde un enfoque de tercera persona, tal como el empleado por la I.A.

Ahora bien, un investigador funcionalista podría perfectamente preguntar: "¿Por qué debiera la I.A. entregar conciencia fenoménica a una máquina?" Podría éste decir que la conciencia fenoménica es irrelevante para la actividad cognitiva, ya que sólo se necesita el algoritmo correcto para capturar toda actividad considerada como inteligente, y que la conciencia es sólo un epifenómeno[37]. Existen varias razones para replicar esa afirmación. Primero, debido a que ésta resulta ser parte de nuestra realidad cognitiva, es decir, es tan real como otros fenómenos mentales que caracterizan el ser humano, y segundo, debido a que participa activamente en conductas que requieren inteligencia[38], por lo que sería equivocado ignorarla debido a su naturaleza problemática. En otras palabras, la conciencia fenoménica juega un rol causal en tareas inteligentes, por lo que negar su importancia debido a la dificultad de su estudio e implementación, es un error. Por ejemplo, considere la ventaja de la experiencia fenoménica del miedo o el estrés al detectar un peligro inminente, y cómo eso posee un rol causal en la conducta al momento de cruzar una calle y ver como los autos vienen acercándose peligrosamente. Además, la experiencia subjetiva del mundo moldea la manera en la cual cada criatura se desenvuelve y percibe su ambiente, y por tanto, forma parte de la manera cómo responderá a su entorno, y es por eso que la I.A. debiera incluir su implementación al momento de diseñarlas. Para graficar la importancia de esto último, modificaremos el conocido experimento mental presentado en la sección 2.3.2 denominado 'Espectro Invertido' (Locke, 1690). Este último resumía el caso de dos personas igualmente funcionales desde un punto de vista psicológico, pero cuyas experiencias fenoménicas de color rojo/verde se encuentran invertidas. El experimento original demuestra que es plausible implementar sistemas funcionalmente idénticos, pero que no explican ni cómo ni por qué se produce la experiencia subjetiva. Ahora bien, modifiquemos el mismo experimento para mostrar cómo la conciencia fenoménica es relevante para la conducta inteligente. Volvamos con nuestros amigos Pedro y Pablo. Supongamos que a ambos se les realiza una intervención quirúrgica de la porción occipital de la corteza, es decir, la zona del cerebro que controla el campo visual. Un día cualquiera, mientras duermen y sin previo aviso, se les realiza la operación sin que ellos se enteren. La cirugía consiste en intervenir sus cerebros de manera tal que, al despertar, Pedro posee la experiencia visual subjetiva de rojo/verde de Pablo, y viceversa. Se debe señalar y tener en consideración que la operación sólo cambia la experiencia fenoménica de los individuos, ya que el resto de los sistemas, estados mentales y organización funcional de éstos, siguen intactos. Al despertar y producto de la intervención, ambos dicen cosas como: "Que raro, las hojas de los árboles son rojas ahora" o "llamen a un doctor, ¡mi sangre se ha vuelto verde!". En ese mismo instante, se les pide que realicen la siguiente tarea: Para sanarse de aquella rara experiencia de cambio de colores, deben elegir entre tomar la pastilla roja y la verde. Las pastillas no tienen ningún tipo de etiqueta ni nombre. Se le dice a cada uno que la roja lo sanará, mientras la otra hará que se sientan aún más

[37] Se desarrolló el concepto de epifenómeno en 1.2.2.

[38] Se desarrolló sobre esto en la sección 3.3.

enfermos. Pero los sujetos se encuentran con sus experiencias fenoménicas alteradas, por lo que, irremediablemente, eligen la pastilla contraria a la que realmente quieren elegir. El hecho que los espectros invertidos de uno y otro estén re-invertidos afecta directamente su decisión, no permitiéndoles llevar a cabo lo que su mente promueve (cuidarse y evitar seguir enfermos). Es decir, los sujetos poseen la misma organización funcional que tenían antes de la operación, y sobre la base a eso, el Funcionalismo entregaría exactamente la misma descripción de la experiencia de Pablo de "veo rojo" que la experiencia de Pedro de "veo rojo", por lo que el Funcionalismo prevería que *no* deberían tener problemas en escoger la pastilla que los hará sentir mejor. Sin embargo, los sujetos escogen mal porque la experiencia subjetiva de color se encuentra alterada en ambos. En conclusión, Pedro y Pablo dependen de sus experiencias personales de rojo/verde para poder desempeñar correctamente una decisión. La experiencia subjetiva de color juega un rol básico e irreemplazable en el correcto funcionamiento cognitivo de los sujetos al momento de realizar ciertas tareas inteligentes, tales como decidir correctamente con respecto a buscar el bienestar de sus organismos.

Es posible, por lo tanto, cuestionarnos sobre la utilización de algoritmos y una disposición funcional adecuada como mecanismos suficientes para entregar una dimensión tan importante de nuestra cognición como la conciencia fenoménica, ya que existe una variedad de problemas filosóficos al momento de intentarlo. El primer problema que la I.A. enfrenta es que no existe algo así como un *what it is like to be* una máquina programada; es decir, no existe una manera de ser máquina. ¿Cómo podemos diseñar máquinas que necesiten conciencia para realizar tareas inteligentes si no sabemos (ni podemos saber) *what it is like to be a machine*? De la misma manera en la que Nagel plantea que nunca podremos saber qué se siente ser un murciélago, no podemos saber qué se siente ser una máquina programada. La utilización de un algoritmo como componente esencial de la arquitectura falla en dar cuenta de este hecho, dado que no hay nada en un algoritmo que lo entregue ni explique por qué se produce. Puedo actuar como una máquina, e implementar los dispositivos de recepción de *input* y operar en función de algoritmos, sin embargo, eso no es suficiente, ya que con ello entendería qué se siente para un humano ser una máquina programada, pero jamás sabría que se siente para una máquina programada ser una máquina programada. Esto se debe a que me encuentro limitado por mi propia mente, por mi propia experiencia subjetiva. He ahí la dificultad de crear una máquina programada basada en algoritmos: Nada en el algoritmo entrega ese "qué se siente" ser una máquina, además de que su utilización no explica ni sugiere por qué los estados mentales son acompañados por una experiencia subjetiva. Adicionalmente, cabe preguntarse sobre la factibilidad de capturar todas las experiencias y tareas que un ser humano pudiera realizar en algoritmos. Piense, por ejemplo, en estados alterados de conciencia, tales como encontrarse embriagado. ¿Cuál es al algoritmo para estar con varias copas de más? Es cierto, sería posible escribir un algoritmo que haga a un robot simular todos los comportamientos asociados a encontrase en un estado así, tal como caminar con dificultad y decirle a sus interlocutores lo mucho que los aprecia. Pero la conciencia fenoménica no consiste sólo en eso, ya que como se ha tratado de explicitar, la replicación de comportamiento no equivale a replicar conciencia. Existe una experiencia privada y personal de encontrarse en un estado así, y nos encontramos "atrapados" en ésta.

Además, cabe preguntarse: Dado que no podemos acceder al *what it is like to be* una máquina programada, ¿Cómo podemos saber cuál es el programa o algoritmo correcto o suficiente para asegurarnos que la máquina posee conciencia? Es decir, el carácter deíctico de ésta nos impide acceder a la naturaleza de la experiencia subjetiva de una máquina, lo que es una dificultad al momento de intentar "capturarla" en un algoritmo.

Lo anterior es posible graficarlo a través de un breve experimento mental. Imagine una persona que posee una enfermedad ficticia denominada 'Ceguera al Color Fulminante', la cual es contraída por el sujeto Luis al momento de nacer. Dada la naturaleza de la grave enfermedad, el sujeto Luis jamás ha tenido la oportunidad de experimentar la vida en colores, y no sabe cómo diferenciar el azul, por ejemplo, del gris o del verde. Supongamos que, con el correr de los años, Luis va a la Universidad y aprende todo lo relativo a la visión en colores, incluyendo teoría del color, Física, Psicología de la percepción, la biología del ojo, cómo el color es procesado por el cerebro, etc. En definitiva, Luis logra un conocimiento total y completo sobre todo lo que se puede saber sobre qué es ver en colores. Al estudiar y aprender toda la información sobre esto, Luis aprende el mismo tipo de información que un programa computacional podría aprender: Frecuencias de ondas lumínicas, comportamiento refractario de la luz, comportamiento de los receptores de *input*, etc. pero Luis nunca aprende qué se siente ver en colores, no puede vivir la experiencia fenoménica de ver el mundo como las personas que si pueden. A su experiencia le falta algo que la información objetiva (aquella información que se le entrega a una máquina programada) no le puede entregar: La experiencia subjetiva de verlos. Lo que hace relevante a esto último es que el aspecto fenoménico de la conciencia no se produce en función de un determinado algoritmo de visión, porque no hay nada en un algoritmo que contemple el hecho de que exista algo así como la experiencia subjetiva de ver en colores. En definitiva, este último punto presenta un problema sin solución para la I. A.: Cómo determinar si una máquina posee conciencia en primer lugar. Es decir, si los planteamientos de Nagel son correctos, ¿Cómo podría la disciplina verificar si la I.A. logró crear una máquina consciente? Si no es posible determinar qué se siente ser una máquina debido a que nos encontramos atrapados en nuestra propia experiencia subjetiva humana, y si se ha demostrado que los criterios conductuales son poco confiables, ¿Qué solución se plantea, por ejemplo, en el trabajo de *Machine Consciousness* (o MC4) para saber si una máquina posee algún grado de conciencia fenoménica? Esto último punto posicionaría un problema sin solución tanto para el Funcionalismo como la I.A. Fuerte.

Existe un segundo problema con el diseño de máquinas conscientes basadas en una premisa funcionalista. Éste radica en un postulado esencial de la teoría, y es el hecho de que la realización física o sustrato de la realización es irrelevante, ya que lo determinante para la actividad cognitiva son las relaciones causales del sistema. Conceptualmente, eso es discutible, y para profundizar en ello, piense en el siguiente experimento mental (siguiendo el argumento de Block 1995a): Imagine que la organización funcional necesaria para la experiencia consciente de oler una rosa fuese implementada a través de las interacciones entre todos los habitantes de China. Suponga que se logra que el gobierno de China lleve a cabo el funcionamiento de una mente humana por una hora. Se le provee a cada uno de sus habitantes de radios para comunicarse entre ellos de la misma manera en la cual las neuronas en un cerebro se conectan, y para comunicarse también con un cuerpo artificial. Los movimientos del cuerpo son controlados por las señales de radio, y las señales se traspasan de acuerdo con instrucciones que los habitantes de China reciben a través del funcionamiento de satélites en el cielo. Las instrucciones poseen la característica de hacer que la gente funcione como las neuronas de un cerebro y las transmisiones de radio como las sinapsis, por lo que estos duplican la organización causal de un cerebro humano normal a la perfección. La pregunta es: ¿Puede este sistema tener experiencias subjetivas, o sentir algo? En otras palabras, una vez que se implementa todo lo que el Funcionalismo establece como necesario para la creación de actividad cognitiva inteligente, ¿Sería posible que la sola organización funcional del sistema permitiera afirmar que la nación China posee la experiencia subjetiva de oler un aroma? Intuitivamente, la respuesta es no. Sería imposible

pensar que la sola organización funcional de un estado mental pueda generar la experiencia subjetiva, y que en definitiva, la nación China no tiene la experiencia cualitativa de oler una rosa. Por supuesto, sería posible afirmar lo mismo con respecto al cerebro: ¿Qué hace suponer que el disparo eléctrico de un grupo de neuronas de lugar a una experiencia subjetiva? Pero las personas que replican de esa manera fallan en ver la real intención del argumento de Block, que es posicionar el hecho de que sería conceptualmente posible que un sistema duplique funcionalmente a un humano normal con respecto a sus estados mentales y aún así, no experimentar ningún tipo de *quale*. En resumen, el experimento de la nación China separa a la Ciencia Cognitiva de la ingeniería, toda vez que otorga preponderancia a la organización causal de un determinado sistema. Sin embargo, por complejo que éste sea, no es razón suficiente para generar una experiencia subjetiva, ya que no es posible caracterizar la experiencia fenoménica en términos funcionales.

Sería posible afirmar entonces que una concepción funcionalista para replicar estados fenoménicos en una máquina no basta para agotar la explicación. Al parecer, se necesitaría entregar la posibilidad a una máquina de ser capaz de identificarse como un punto de vista al cual le pasan cosas, en otras palabras, un sujeto que tiene determinado estado fenoménico sobre sí mismo y el mundo. La experiencia subjetiva del mundo, además, forma parte esencial de nuestra experiencia y nuestra constitución de *self*. Lo anterior nos lleva a una tercera observación: La conciencia fenoménica le ocurre a un punto de vista. ¿Cómo puede la arquitectura basada en algoritmos entregar la sensación de individualidad necesaria para experimentar el mundo? Para que esa sensación de que soy “yo” el que se encuentra estresado en estos momentos, existe una parte fundamental: La sensación de que hay un sujeto al cual le ocurre dicha experiencia. Y nada en un algoritmo entrega eso. Existe un sujeto de la conciencia, un individuo que debe en ciertas circunstancias identificarse a sí mismo y los estados en los que se encuentra, sin necesidad de que una observación externa lo confirme. Pero nada en la sucesión de pasos y estados del algoritmo sugiere cómo se produce, ni cuál sería su función, ni que ese punto de vista sea importante para la actividad inteligente. Suponga, por ejemplo, que en algún país del mundo se encuentra la tecnología capaz de remover la conciencia fenoménica. Suponga que a los filósofos eliminativistas o reduccionistas se les ofrece la posibilidad de someterse a ésta como sujetos experimentales. La pregunta es: ¿Se sometería gente como Churchland y Dennett a este tipo de operación? A título personal, rechazaría tal intervención. Al someterme a este tipo de intervención, estaría perdiendo algo que forma parte de mi cotidianidad, que me permite navegar el mundo de la manera que lo hago, que se relaciona con mi sensación corporal y perceptual, ya que la experiencia fenoménica constituye parte de mi identidad como individuo, y por lo tanto, la remoción de ésta, afectaría directamente mi vida mental. Desde una óptica distinta, imagine qué ocurriría si fuera posible implementar una operación que permitiera eliminar la experiencia subjetiva del dolor en pacientes que sufren jaquecas o se encuentran en estados terminales de alguna enfermedad dolorosa como la artritis. Si la dimensión fenoménica del dolor es sólo la activación de neuronas y, en realidad, es sólo un espejismo, no explicaría el hecho de que la mayoría de esas personas (quizás todas) se someterían gustosos a tal intervención. Es que la dimensión fenoménica forma parte de la constitución de un punto de vista, y es posible sostener que se relaciona con la construcción de nuestro *self*.

Es por eso que autores interesados en el estudio del *self* como Damasio (2000) sugieren una conexión entre la conciencia fenoménica y el desarrollo de una identidad propia o *self* para llevar a cabo determinadas conductas, y sostiene lo siguiente:

Si 'conciencia de sí mismo' equivale a 'conciencia con sensación de self', la expresión abarca toda la conciencia humana [...]. Agregaría que el estado biológico que describimos como sensación de self y la maquinaria biológica responsable de su génesis pueden desempeñar un papel en la optimización del procesamiento de los objetos por conocer: Quizás la sensación de self no sólo sea necesaria para conocer en sentido propio, tal vez pueda influir en el procesamiento de cualquier cosa por conocer [...]. Cuando interpelo el problema del self, interrogo el tema de los Qualia con relación a la representación del organismo que posee conciencia. (pp. 35-36.)

En otras palabras, el autor sostiene que al investigar sobre la noción de *self*, también se investiga sobre cómo un individuo se plantea ante el mundo y genera un conocimiento de éste. Es que la conciencia fenoménica nos permite la función biológica crítica que hace posible conocer la pena o la alegría, el sufrimiento o el placer, el orgullo o la vergüenza, aspectos de la vida que permiten desarrollarnos apropiadamente en el contexto cultural/ social al cual pertenecemos.

Lo anterior nos da la oportunidad de aventurar una teoría que bien podría constituir una tesis en sí misma: La conciencia fenoménica forma parte fundamental de procesos cognitivos tales como la empatía o emociones que permiten "ponerse en el lugar de otro", lo que facilita la creación de sociedades y acuerdos[39]. El reciente trabajo llevado a cabo en relación con el vínculo entre 'Neuronas Espejo[40]' y el rol que éstas desempeñarían con respecto a las capacidades cognitivas ligadas a la vida social, podría encontrarse relacionado a la conciencia fenoménica, ya que el acceso a saber qué se siente doler o alegrase formaría parte de este proceso de ponerse en el lugar de un otro. Sin ir más lejos, investigación en esta área podría llegar a determinar algo así como '*Qualia* culturales', tales como las sensaciones asociadas a ser pecador, culpable, o vencedor. Más investigación además de la incorporación de otros marcos teóricos sería necesaria para profundizar en este punto, por lo que el desarrollo en esta materia es menester de otro trabajo de tesis. Sin embargo, se presenta la hipótesis recién señalada como ejemplo de que la conciencia fenoménica forma parte importante en la creación de *self*, haciendo manifiesta una nueva función de ésta dentro de la actividad inteligente humana.

En resumen, una máquina consciente capaz de sentir el mundo, posee ciertas características que le permitirían ir más allá del algoritmo que la determina, es decir, la conducta inteligente podría verse asociada no sólo al correcto funcionamiento de un grupo de reglas bien establecidas y bien formuladas, sino también a la interacción de esta capacidad cognitiva consciente con el resto del sistema cognitivo, lo que permitiría, entre otras cosas, alterar dichas reglas en favor de una mejor respuesta al ambiente en el cual se encuentra contextualizado. Y no sólo eso, podría contribuir al mejor desarrollo de máquinas capaces de exhibir algún grado de autosuficiencia. Una de las capacidades esenciales de la actividad cognitiva inteligente es la capacidad para adaptarse y modificar conductas anteriores o previamente establecidas ante la eventualidad de situaciones novedosas. Es

[39] Robbins (2008) también refiere a este punto, particularmente con respecto a lo del 'Dolor Social' mencionado en la sección 3.3.

[40] Se denomina Neuronas Espejo a una cierta clase de neuronas en el área de Broca y corteza parietal humana que se activan cuando una persona desarrolla la misma actividad que otro ser humano realiza, es decir, las neuronas del individuo reflejan como un espejo la acción de otro. Se ha llegado a considerar que las Neuronas Espejo son uno de los descubrimientos más importantes de las Neurociencias en la última década, dada la posibilidad de entender aspectos sociales desde la Neurociencia, realizando un nexo claro entre Ciencias Sociales y Ciencias Naturales.

por eso que ciertos autores en I.A. (Alexander 2007, Chella y Manzotti 2007) aseveran que el estudio y comprensión de la conciencia pudieran otorgar un mejor cimiento para el control de la conducta en situaciones en las cuales se requiera de autonomía por parte de una máquina. Es en estas circunstancias que aproximaciones tales como la 'Cognición Situada' y la 'Cognición Corporalizada'[41] pudieran ser capaces de arrojar luz sobre el problema de la conciencia, toda vez que son capaces de concebir la cognición y la conducta inteligente considerando un agente situado en un contexto tanto físico como social, y que se encuentra en contacto con un ambiente real y lleno de particularidades, en vez de una concepción abstracta generada por un procesador para propósitos generales. En ningún caso se sugieren como paradigmas que compitan o sean mutuamente excluyentes (no en relación con la conciencia fenoménica al menos), sino más bien como una aproximación que pudiera ofrecer consideraciones capaces de entregar nuevas perspectivas que relacionen conciencia con un punto de vista y el contexto en el cual ocurre.

El presente trabajo también nos lleva a realizar algunas consideraciones sobre algunos aspectos de la conciencia desarrollados en el capítulo 3. Se ha discutido la dificultad de estudiar la conciencia desde la tercera persona. La experiencia cualitativa no puede ser medida por pruebas, dada su característica de tener una ontología de primera persona. Pero, ¿Por qué debiera ser así? ¿Por qué asumir que todos los hechos del mundo son igualmente accesibles para el observador estándar de tercera persona? El hecho de que no pueda ser observable desde la tercera persona no es razón suficiente para negar su realidad psicológica. El estudio de la conciencia fenoménica es el estudio de las apariencias, por lo que es sensato no reducir su explicación a la búsqueda de un correlato neuronal o de un dispositivo que la reproduzca debido a que estos tienen un carácter "objetivo". Es al momento de negar el aspecto subjetivo de la conciencia o reducirla a algo objetivamente observable que se produce un error, y es una posición que admita, primero, el hecho de que existe, y segundo, que cumple una función en la conducta inteligente, la que guiará el trabajo en I.A. de manera adecuada para otorgar dicha facultad a un dispositivo artificial.

La conciencia fenoménica, ontológicamente, puede existir independientemente de la conducta, algún rol funcional o sus relaciones causales. Epistemológicamente, gracias a la conducta sabemos que otras personas tienen estados mentales conscientes. Causalmente, la conciencia se relaciona con los *input* y *output* de nuestro organismo. Sin embargo, parece plausible concebir máquinas que posean una conducta asociada a la actividad consciente sin tener experiencias cualitativas (piense en un robot *zombie* que sea capaz de interactuar normalmente con su ambiente al igual que un humano, pero que no presente experiencias sensoriales subjetivas), así como también es posible que una máquina sea capaz de tener esas experiencias sin que sean transmitidas ni utilizadas y que no sea posible saber que las tiene. Lo anterior lleva a sostener algo así como un "Principio de Independencia" de la conciencia con respecto a la conducta observable y la función que esta pueda desempeñar.

Al tratar de explicar por qué algunas posiciones buscan ignorar aquello que la teoría no puede explicar (en este caso, el carácter cualitativo de las experiencias mentales conscientes), es posible aventurar un prejuicio teórico. Este prejuicio consistiría en tratar de dar cuenta del mundo en términos completamente materialistas, lo que a su vez produce un rechazo a que exista algún tipo de aspecto de la realidad que no pueda ser reducido a un concepto físico. Sin embargo, es posible identificar fenómenos mentales tales como la conciencia que no pueden ser reducidos a lo material, lo que nos obliga a cuestionar si los métodos y premisas que usamos son los correctos, y nos da la oportunidad de constantemente pensar la mente desde distintas aproximaciones que puedan arrojar luz

[41] Ver sección 3.3 para una breve descripción de ambas.

sobre los problemas que el Materialismo (u otras visiones de la mente y de la realidad) pueda presentar.

No obstante, llama la atención que no exista una necesidad de reducir otro tipo de fenómenos a lo material o conductual aparte de los mentales. Nadie se atrevería a definir la existencia de los ojos única y exclusivamente en términos funcionales de desempeño de conos y bastones. Empero, existe una necesidad de reducir nuestros dolores y sensaciones a algo físico. Probablemente, esto se debe a que fenómenos tales como la conciencia fenoménica, el miedo o el placer, poseen un componente subjetivo, privado, al que únicamente accedemos desde la introspección, y no sólo eso, tampoco puede ser definida ostensiblemente. Existe una necesidad de aclarar y develar los misterios de cómo las cosas parecieran ser, de eliminar las apariencias para indagar en la real naturaleza de los fenómenos, y sin lugar a dudas, esa ha sido la función de las ciencias. Pero, ¿Qué ocurre cuando la apariencia *es* la realidad? Es decir, qué hacer cuando nos enfrentamos a fenómenos mentales - tales como la conciencia fenoménica - en los cuales la apariencia de una corriente de pensamientos es la realidad, nuestra realidad, concreta y objetiva. ¿Por qué rehusarse a la idea de que algo así existe, incluso si la teoría funcionalista falla en entender o incorporar dicho aspecto en su explicación de la vida mental? Más aún, existe una tendencia a argumentar en contra de la conciencia a través del clásico ejemplo del agua y H2O. Se dice que la conciencia, en determinado momento, será reducida a los estados del cerebro en la medida que la Neurociencia avance y genere una teoría que sea capaz de aprehender los preceptos de la conciencia fenoménica. Es decir, de la misma manera en la que la teoría avanzó y logro reducir el concepto de “agua” a través de una teoría explicativamente más poderosa, de la misma manera la conciencia será reducida a los estados del cerebro. Ahora bien, la analogía del agua/H2O y conciencia fenoménica/ estados del cerebro presenta una falla, que consiste precisamente en lo que se ha venido comentado hasta acá: Lo que hace particular a la conciencia es que la apariencia, o cómo parecen ser las cosas, es la realidad, por lo que no se debe buscar una reducción de un fenómeno “aparente”. Es precisamente este último rasgo, “la apariencia de una realidad subjetiva que le ocurre a un individuo”, la que se debe explicar e implementar, ya que es un dato real. Y esto último no sucede con el agua, en el cual simplemente se transforma un concepto en otro que lo contiene, explicativamente hablando. Por lo tanto, una explicación de la conciencia debe hacerse cargo de este rasgo cualitativo concreto, y no buscar una explicación que lo ignore.

Muchas de las confusiones sobre la conciencia podrían ser resueltas al considerar una simple aclaración con respecto a la relación entre un fenómeno subjetivo y la objetividad de la ciencia. Se dice que la ciencia es por definición, objetiva. Se sigue de lo anterior que no puede haber una ciencia de la conciencia, ya que no se sabe cómo estudiar objetivamente lo subjetivo. Sin embargo, existe una confusión con respecto a qué se habla cuando se utilizan los términos “subjetivo” y “objetivo”. Según Searle (1998), se necesita hacer una distinción entre lo que es ontológicamente objetivo y ontológicamente subjetivo, y lo que es epistemológicamente objetivo y epistemológicamente subjetivo. A continuación se hará una breve distinción entre ellos:

A.- Afirmación epistemológicamente objetiva. Ejemplo: “El Instituto Nacional fue fundado en 1813”. Son aquellas afirmaciones que son objetivamente verificables, es decir, que su verdad o falsedad no dependen del observador.

B.- Afirmación epistemológicamente subjetiva. Ejemplo: “La cerveza Cristal es mejor que la cerveza Escudo”. Difiere de la anterior en que su verificación si depende del observador. En este caso, la verdad o falsedad del hecho depende de la opinión de éste,

debido a que no existe ningún prueba o mecanismo objetivo que permita establecer que lo referido por la proposición se corresponda con una situación en el mundo.

C.- Modo de existencia subjetivo. Ejemplo: Mi sensación de tensión en los hombros luego de llevar escribiendo 5 horas en el computador. Mi sensación de cansancio en mis hombros es ontológicamente subjetiva, ya que existe en la medida que es experimentado por un sujeto, en este caso, mi persona. Por extensión, todos los estados conscientes son ontológicamente subjetivos.

D.- Modo de existencia objetivo. Ejemplo: Montañas, células embrionales, arena. Estas entidades tienen una ontología objetiva, ya que no deben ser experimentados por un humano o animal para que existan.

Hecha la distinción, según el autor, la objeción es ambigua. La Ciencia es objetiva epistemológicamente, ya que busca datos e información que sea independiente del investigador que la busca, y la cantidad de electrones de un átomo de hidrógeno será siempre la misma, independientemente de quién sea el científico. (Por extensión, la afirmación de que la cerveza Cristal es mejor que la cerveza Escudo no es una afirmación científica). Pero el *quid* del asunto es éste: Que la ciencia busque objetividad epistemológicamente hablando, no elimina el hecho de que existan entidades ontológicamente subjetivas que pueden ser objeto de estudio científico, es decir, ser tratadas igualmente a las entidades que son ontológicamente objetivas. En otras palabras, puede haber conocimiento epistemológicamente objetivo sobre entidades que son ontológicamente subjetivas. En conclusión, dada la objetividad de la Ciencia, no es impedimento plantear una Ciencia epistemológicamente objetiva sobre un dominio ontológicamente subjetivo, por lo que considerar que el carácter subjetivo de la conciencia es un impedimento para un estudio científico de esta, es un error.

Pasando a otro aspecto de la presente discusión, cabe detenernos un instante sobre el trabajo en I.A. en relación con la dicotomía *A/P-consciousness*.

El campo de *Machine Consciousness* no es un campo unificado de investigación que tenga metas claras, y eso es lo que explica que determinadas críticas y diálogos con otras disciplinas puedan, en ciertas ocasiones, verse entrampados en aclaraciones sobre cuál es la verdadera naturaleza del objeto a ser implementado. He ahí la pertinencia de incluir las definiciones de Block y Nagel. Los planteamientos de ellos con respecto a la conciencia fenoménica se encuentran asociados a lo que en la sección 3.3 denominamos como MC4, mientras que el trabajo en MC2 y MC3 se encuentra asociado a lo que el mismo Block denomina *A-consciousness*. Por otro lado, nociones como Test de Turing se conectan particularmente a MC1. Desde hace pocos años, la conciencia ha dejado de ser una noción peligrosa para la comunidad que trabaja en I.A. En los seres humanos, ésta corresponde a una colección de distintas características de la cognición. Los investigadores en I.A. han comenzado a interesarse en ésta cada vez más, de manera tal de, inicialmente, entenderla, para luego intentar replicarla en agentes. Sería posible decir que existe consenso sobre la distinción entre los aspectos fenoménicos y los cognitivo-funcionales, reflejando la distinción realizada por Block (1995b) entre *P-consciousness* y *A-consciousness*, respectivamente. El trabajo en el primer tipo de conciencia corresponde a lo que se conoce como el 'Problema Difícil de la Conciencia' (sección 3.1), y el trabajo en ésta requiere compromisos teóricos fuertes, ya sea en favor de una posición extrema que permite sostener declaraciones tales como que "los termostatos, efectivamente, poseen estados fenoménicos", o bien enfocando el estudio en los aspectos cognitivos desde una posición más moderada que sugiera que el trabajo en I.A. se relaciona con una simulación más que una replicación. Sin embargo, es probable que el trabajo en estas áreas muchas veces se

vea entrampado por la dicotomía, toda vez que en el trabajo de la I.A. los planteamientos filosóficos que residen bajo los campos experimentales que buscan implementar *A-consciousness* o *P-consciousness*. Pero ninguna busca implementar ambas. Si el objetivo central de la I.A. es replicar las capacidades cognitivas humanas, sería óptimo implementar una aproximación que considere a ambas concepciones de la conciencia de manera simultánea, es decir, buscar una arquitectura, teoría o implementación que pueda entregar explicaciones y modelamientos de ambos aspectos de la conciencia. Si es que realmente se quiere lograr una teoría de la conciencia, esta debería incluir una explicación sobre ambos aspectos de la dicotomía, ya que privilegiar una u otra, irremediablemente deja algo afuera de la explicación.

En la actualidad, el trabajo en máquinas conscientes proyecta muchos beneficios, tales como avanzar en el desarrollo de un correlato neuronal en humanos, mejorar los diagnósticos sobre fenómenos mentales involucrados en pacientes en coma, o el desarrollo de extensiones prostéticas para recuperar funciones visuales o auditivas (en Gámez, 2007). Lo anterior, permite referir a los potenciales problemas éticos futuros que traería consigo la creación de máquinas conscientes. Es decir, existe la posibilidad de que éstas dejen de ser consideradas como herramientas de investigación, sino como co-participes de la sociedad en la que vivimos. Sería interesante, por ejemplo, cuestionarnos sobre el hecho concreto de experimentar con máquinas que dan señales claras de exhibir algo así como "dolor". Experimentar con máquinas se convertiría un tópico tan controversial como la experimentación con animales o el trabajo de clonación que se desarrolla en la actualidad, desde un punto de vista ético así como legal, ya que, de ser posible implementar entidades conscientes, aquello tendría un impacto sobre los aspectos legales de experimentar con entidades susceptibles de sufrir o tener voluntad propia. Es más, piense en las dificultades legales que podrían presentar un caso en el cual un autómata dotado de voluntad no desarrolle las funciones para las cuales fue diseñado: La responsabilidad de su desempeño no dependería de la programación entregada por una determinada compañía, sino por las propias decisiones de la criatura en cuestión. Lo anterior, incluso, podría suponer un "problema" *a priori* con respecto a crear máquinas que exhiban conciencia fenoménica. Sin embargo, es un argumento que no se relaciona con los fines de los distintos programas de investigación que ven en las maquinas conscientes los beneficios de mejorar el desempeño y construcción de máquinas, relegando esa discusión, momentáneamente, a un segundo plano.

En definitiva, el cruce en el trabajo entre Filosofía de la Mente, Psicología e Inteligencia Artificial ha encontrado un punto de discusión y diálogo encuentro (y quizás de desencuentro) que pocos temas plantean. Es que la conciencia, aquella capacidad única y personal en la cual poseemos como individuos, es la que se ha erguido como una problemática distintiva a otras - tanto para la Ingeniería como para la Filosofía de la Mente - debido a su aspecto cualitativo. El problema de la implementación de conciencia en artefactos creados por el hombre es un campo investigativo relativamente reciente, pero que trajo consigo la necesidad de comenzar a revisar si la solución Materialista/ Funcionalista de la mente es, efectivamente, la teoría más adecuada para comprender los fenómenos mentales. Es menester de estas disciplinas entregar una propuesta que incluya esta dimensión de la cognición en el diseño de una máquina, si es que efectivamente se busca no dejar de lado aspectos fundamentales de los fenómenos mentales y la actividad inteligente en máquinas, abriendo la oportunidad de entregar una dimensión psicológica a esta modelación, toda vez que la actividad inteligente le ocurre a un punto de vista, a la "Presencia de un Alguien" (Damasio, 2000) que se desenvuelve en el mundo y su ambiente. Es el diálogo interdisciplinario el encargado de comenzar nuevas discusiones y revisar

algunas que se habían considerado zanjadas en pos de un mejor entendimiento de qué es lo que acontece en esta compleja dimensión cognitiva que llamamos vida mental.

Bibliografía

Alexander, I. (2007): “Machine Consiousness.” En Schneider and Velman (eds.) The Blackwell Companion to Consciousness. Blackwell Publishing, Mass., pp. 87-98.

Baars, B. (1993): *A Cognitive Theory of Consciousness*. Cambridge University Press.

Bechtel, W. et al. (1998): “The Life of Cognitive Science.” En Bechtel y Graham (eds.) *A Companion to Cognitive Science*. Malden, Mass., Blackwell, pp. 1-104.

Blackmore, S. (2005a): *Consciousness: a Very Short Introduction*. Oxford University Press.

_____. (2005b): *Conversations on Consciousness*. Oxford University Press.

Block, N. (1995a): “The Mind as Software of the Brain.” En Smith y Osherson (eds.) *An Invitation to Cognitive Science,* 2nd Edition, Vol. 3: Thinking.Cambridge, Mass., MIT Press, pp. 377-426.

_____. (1995b): “On a Confusion about a Function of Consciousness.” En Block et al. (eds.) *The Nature of Consciousness*. Cambridge, Mass., MIT Press, pp. 375-415.

Chalmers, D. (1996): *The Conscious Mind: In Search of a Fundamental Theory.*Oxford University Press.

Chella, A. y Manzotti, R. (2007): “Artificial Intelligence and Consciousness.” En *AI and Consciousness: Theoretical Foundations and Current Approaches: Papers from the AAAI Fall Symposium*,9-11 de Noviembre, 2007, Arlington, Virginia, USA, pp. 1-8.

Churchland, P. (1984): *Matter and Consciousness*. Cambridge, Mass., MIT Press.

Copeland, J. (1993): *Artificial Intelligence: A Philosophical Introduction.* Malden, Mass., Blackwell Publishers Inc.

Cummins, R. (1983): *The Nature of Psychological Explanation.* Cambridge, Mass., MIT Press / Bradford Books.

Damasio, A. (2000): *Sentir lo que Sucede*. Santiago de Chile, Andrés Bello.

Dennett, D. (1996): *Kinds of Minds.* New York, Basic Books.

Flanagan, O. (1997): “Prospects for a Unified Theory of Consciousness.” En Block et al. (eds.) *The Nature of Consciousness*. Cambridge, Mass., MIT Press, pp. 97-110.

Gámez, D. (2007): “Progress in Machine Consciousness.” *Consciousness and Cognition* Volumen 17, 3, pp. 887-910.

Hillis, D. (1998): “Can a Machine be Conscious?.” En Hameroff et al. (eds.) *Toward a Science of Consciousness II.* Cambridge, Mass., MIT Press, pp. 181-84.

Jones, A. (2004): “Five 1951 BBC Broadcasts on Automatic Calculating Machines”. *IEEE Annals of the History of Computing,* 26(2), pp. 3-15.

Locke, J. [1960](1975). *An Essay Concerning Human Understanding*. Oxford, Oxford University Press.

Nagel, T. (1974): "What is It Like to Be a Bat?." En Haugeland (ed.) *Mind Design II.* Cambridge, Mass., MIT Press, pp. 519-28.

Newell, A. (1980): "Physical Symbol Systems." *Cognitive Science,* 4, pp. 135-83.

Nida-Rümelin, M. (1999): "Pseudonormalvision and Color Qualia." En Hameroff, Kaszniak, and Chalmers (eds.) *Toward a Science of Consciousness III The Third Tucson Discussions and Debates.* Cambridge, Mass., MIT Press, pp. 75-84.

Putnam, H. (1968): "Brains and Behavior." En Block (ed.) *Readings in Philosophy of Psychology,* Volumen 1*.* Harvard University Press, pp. 24-36.

Robbins, P. (2008): "Consciousness and the Social Mind." *Cognitive Systems Research,* Volumen 9*,* pp. 15-23.

Rumelhart, D. (1989): "The Architecture of Mind: A Connectionist Approach." En Haugeland (ed.) *Mind Design II.* Cambridge, Mass., MIT Press, pp. 205-32.

Searle, J. (1980): "Minds, Brains and Programs" *Behavioral and Brains Sciences,* 3, pp. 417-59.

_____. (1998): "How to Study Consciousness Scientifically". *Philosophical Transactions: Biological Sciences*, Volumen 353, 1377, pp. 1935-42.

_____. (2004): *Mind: A Brief Introduction*. New York, Oxford University Press.

Smolensky, P. (1989): "Connectionist Modeling: Neural Computational/ Mental Connections." En Haugeland (ed.) *Mind Design II*. Cambridge, Mass., MIT press, pp. 233-50.

Stillings, N. et al. (1995): *Cognitive Science. An Introduction*. Cambridge, Mass., MIT Press.

Turing, A. (1948): "Intelligent Machinery." En Evans and Robertson (eds.) *Cybernetics: Key Papers.* Baltimore, University Park Press, pp. 26-54.

_____. (1950): "Computer Machinery and Intelligence." En Haugeland (ed.) *Mind Design II.* Cambridge, Mass., MIT Press, pp. 29-56.

Apéndice.

En la introducción a este trabajo se mencionó el ejemplo de *Deep Blue* como un caso "exitoso" de creación de una máquina que desempeña actividad inteligente. La palabra en cuestión va entre comillas, debido a que la naturaleza de este éxito depende de lo que se considere como inteligente, pero por sobre todo, debido al nivel de comprensión que *Deep Blue* realmente exhibe sobre las tareas que realiza. ¿Comprende realmente la máquina lo que hace? Esta pregunta no se relaciona directamente con *A-consciousness*, sino con *P-consciousness*. Y es pertinente referir al origen de esta confusión y al problema detrás de éste, el 'Problema de la Intencionalidad'.

El 'Problema de la Intencionalidad' es ampliamente discutido por Searle (1980, 2004). La intencionalidad corresponde a la característica de ciertos estados mentales y eventos que consisten o son acerca de algo. Desde otro punto de vista, se puede decir que gracias a la intencionalidad, un sujeto es capaz de conocer la realidad que lo rodea. Para Searle, sólo los estados mentales pueden ser intencionales, pero no todos, ya que existen otros tales como los estados fenoménicos - por ejemplo, el dolor - que no necesariamente son "acerca de algo". Además, el autor no considera factible dar esas propiedades a una máquina cuya operación esté definida solamente en términos de procesos informáticos realizados sobre elementos definidos formalmente. Las manipulaciones de símbolos formales por sí mismas no tienen ninguna intencionalidad, es decir, son no significativas; no son ni siquiera manipulaciones de *símbolos*, puesto que los símbolos no simbolizan nada. En la jerga de los lingüistas, tienen sólo sintaxis, pero no semántica. La intencionalidad que los programas de computación parecen tener corresponde solamente a las mentes de los que los escriben y usan ese programa, los que les dan los *inputs* e interpretan sus outputs.

Con respecto a la mente, se debe hacer una distinción entre 'Intencionalidad Original' e 'Intencionalidad Derivada'. Por ejemplo, yo tengo en mi cabeza información sobre cómo llegar al Palacio de Gobierno desde un lugar de Santiago. A esta información y estas creencias se les conocen como 'Intencionalidad Original'. El mapa al cual puedo recurrir en internet también posee esta información, posee símbolos y expresiones que refieren y representan una ruta de navegación. Pero el sentido en el cual el mapa posee esta información se deriva de la intencionalidad original de las personas que hicieron los mapas. Por lo que la intencionalidad del mapa fue impuesta por la intencionalidad original de los diseñadores. De la misma manera, la intencionalidad sobre el juego de *Deep Blue* fue impuesta por la intencionalidad original de los programadores humanos. En otras palabras, *Deep Blue* proyecta y utiliza su programación para desplegar las capacidades de los programadores y jugadores de ajedrez que ayudaron a entregarle una serie de pasos finitos que le permitiera llegar a un estado final etiquetado en el programa como el estado 'Jaque Mate', lo que no necesariamente significa que la máquina entiende el juego de la misma manera que un jugador humano lo haría. Sería posible desarrollar más sobre el problema de la Intencionalidad y sus implicancias en el modelamiento de máquinas programadas, pero aquello sería materia de otra tesis y de futura investigación.