



Casa abierta al tiempo

UNIVERSIDAD AUTÓNOMA METROPOLITANA
Unidad Iztapalapa

División de Ciencias Sociales y Humanidades

Posgrado en Humanidades

“¿Habría consciencia en máquinas?”

TESIS

Que para obtener el grado de

MAESTRO EN HUMANIDADES

Línea académica de Filosofía de la Ciencia

PRESENTA:

José Carlos Rojas Reyes

2223802320

reyesjcrr@hotmail.com

DIRECTOR DE TESIS:

Dr. Jorge Martínez Contreras

JURADO:

Presidente:

Dr. Jorge Martínez Contreras

Secretaria:

Dra. Paola Hernández Chávez

Vocal:

Mtro. Juan Francisco Ortiz Moreno

Iztapalapa, Ciudad de México a 28 de mayo de 2025

Índice

Introducción	2
Capítulo 1	5
1.1 Ontología de la consciencia	5
1.2 Otras formas de la consciencia	11
1.3 Ser ese organismo y no otro	13
1.4 Consciencia emergente.....	15
Capítulo 2	18
2.1 Representaciones mentales	18
2.2 Materialismo.....	22
2.3 Epifenomenalismo.....	25
2.4 Interaccionsimo.....	28
2.5 The Hard Problem of Consciousness.....	32
2.6 Reduccionismo	36
Capítulo 3	41
3.1 Otras mentes. Similitud física.....	41
3.2 Similitud conductual	44
3.3 Una hipótesis auxiliar	47
Capítulo 4	50
4.1 La consciencia en seres artificiales	50
4.2 A partir de otros materiales	54
4.3 Todo ya está programado.....	60
4.4 No se comportan como nosotros.....	67
Conclusiones	73
Referencias.....	77

Introducción

Cuando observamos cómo funcionan los cuerpos de los seres vivos notamos que estos poseen variadas características que les permiten realizar diversas funciones. Por ejemplo, pensemos en algo tan común como el pico de una gaviota; su forma larga, delgada y ligeramente curvada en la punta, le facilita la caza de peces y de otras criaturas. Además, sus glándulas de sal ubicadas cerca de la base del pico le permiten eliminar los excesos salinos al beber agua del mar. De igual modo, las gaviotas no solo capturan peces con sus picos, también escarban en la arena, avientan conchas para abrirlas o picotean a los intrusos para defender su territorio.

Tanto los cuerpos de las gaviotas como de otras criaturas presentan sistemas complejos, entrelazados ecológicamente con otros sistemas y que interaccionan de cierto modo dependiendo de las condiciones ambientales.

Por otro lado, describir a las criaturas de este modo tan mecánico probablemente nos haga pensar que podríamos agotar su naturaleza mediante explicaciones detalladas sobre su funcionalidad, materialidad y composición pero, al parecer, dejamos de lado un aspecto esencial de ellas: las experiencias conjuntas con múltiples aspectos funcionales. Esa subjetividad, esa sintiencia que, si bien no la experimentamos en primera persona sí la damos por hecho como si hubiera algo como ser una gaviota o un cocodrilo calentándose bajo el sol.

De acuerdo con ciertos criterios, juzgamos desde la humanidad quienes son conscientes. O sea, observamos otros cuerpos y, a través de un proceso claroscuro, establecemos en quiénes sí se presenta la sintiencia y en quiénes no. Sin embargo ¿es este juicio en realidad preciso y confiable? ¿Siempre utilizamos imparcialmente un mecanismo de atribución de propiedades fenoménicas? O, quizá, en vista de los debates sobre la comunicación y la consciencia en animales, plantas u hongos, nuestra atribución resulte sesgada o errónea.

Ahora, el tema principal de este trabajo consiste en relacionar estas preguntas con la sintiencia de máquinas. Ante el exponencial desarrollo tecnológico creemos que estas preguntas son filosóficamente relevantes y, por ello, queremos aportar en esta ICR una investigación sobre el estado de la cuestión. ¿Pueden las máquinas ser entes conscientes? Si no lo son ahora, ¿podrán serlo en el futuro? ¿En qué sentido podemos entender el concepto

de consciencia aplicado en máquinas? Si estas poseyeran consciencia, ¿cómo serían sus experiencias?

Durante las últimas décadas se han desarrollado diversos proyectos de *inteligencia artificial*, los cuales, en grandes rasgos, tienen como objetivo la creación de máquinas capaces de realizar actividades que normalmente requerirían de la inteligencia humana.

Considerando el desarrollo continuo de estas y de otras máquinas que devienen cada vez más complejas, potentes y sofisticadas, mi interés por la atribución de consciencia no se centra únicamente en los seres vivos, sino, en aquellos artefactos construidos a partir de materiales no biológicos, capaces de ensamblar autos, participar en conversaciones, analizar cantidades colosales de información o vencer ampliamente a los humanos en partidas tanto de ajedrez como en otros juegos de computadora. En otras palabras, como consecuencia de la creciente innovación en máquinas, me pregunto si estas son o podrán devenir sintientes, poseedoras de esta “película privada”, de estas experiencias cotidianas.

Si bien para Descartes los animales no humanos podrían ser especies de máquinas, ponemos en discusión la fuerza instintiva que produce en ellos una reacción ante una acción del medio o cenestésica, considerando si este comportamiento se asemeja a una intencionalidad o a un automatismo total. De cualquier modo, al hablar de máquinas, por muy complejas que sean, debemos asumir que estas no son entes biológicos.

Debo señalar que, a pesar de sumergirnos en un tópico relativamente nuevo y en muchos aspectos poco conocido, la comunidad de investigadores inmersos en estos temas aumenta progresivamente de modo que se constituye una *filosofía de la inteligencia artificial*.

De igual modo, reconozco que la cuestión ha sido abordada por diversos autores en diferentes épocas y que, a pesar de mi curiosidad, jamás podré agotarlos a todos. Por ello, mi selección de textos, aunque se compone en su mayoría por artículos de las últimas tres décadas, no se limita a ellos, pues existen aportes más antiguos.

Por ello, para tratar de ofrecer una contribución relevante, considero primordial comenzar con un análisis conceptual de la consciencia, de sus acepciones ontológicas, neurológicas y fenoménicas. ¿Cómo distinguimos a la consciencia? ¿Qué contenidos posee? ¿Cómo surge? ¿Cuáles son sus correlatos neuronales?

En filosofía tradicionalmente la consciencia se encuentra asociada con los conceptos de intencionalidad; en la *fenomenología*, incluso, con los de autoconsciencia o reflexión. En

la mayoría o totalidad de las teorías filosóficas tradicionales se trata de una sensación interna, de *qualias*.

Una vez que, a través de nuestra inmersión en los problemas cardinales dejemos claro nuestro concepto de consciencia, continuaré con una materia crucial referente a la manifestación de la sintiencia mediante las *representaciones mentales*. Si bien la posesión de estas nos aseguraría la posesión de algún tipo de sintiencia, es menester preguntarnos qué tipo de entidades son y cómo se relacionan con el “mundo”, lo cual, inevitablemente deriva en una crítica del dualismo.

De igual manera debemos tener en mente que el *Hard Problem of Consciousness* (HPC) nos marca un punto de partida para dudar por qué tenemos esta sintiencia adherida y no somos simplemente objetos sin propiedades fenoménicas.

También, es crucial mantener una postura abierta de modo que, para profundizar en el terreno de la consciencia no podemos limitarnos a uno o pocos campos de estudio; la multidisciplinariedad que rodea el tema nos obliga a tomar y reconocer los aportes desde las diferentes perspectivas de las diversas investigaciones dedicadas a ello.

Así, una vez superada la composición de los dos primeros capítulos, el tercer subtema se centrará en nuestras evidencias para la atribución de agencia mental. ¿Qué soporta nuestras creencias al adscribir propiedades mentales? ¿Qué razones damos para convencernos de que otras entidades son conscientes?

Finalmente, como resultado de la previa investigación entraremos en lo que concierne a la sintiencia de máquinas, de manera que, al no poder ofrecer una prueba directa de su agencia o no agencia mental, me daré a la tarea de examinar las objeciones en contra de suponer su consciencia para así comprender cómo y en qué medida estas son o pudiesen ser conscientes.

Capítulo 1

1.1 Ontología de la consciencia

Donde sea que observemos encontramos entidades dinámicas, objetos en movimiento cuya naturaleza intentamos revelar mediante las explicaciones que nos damos, explicaciones que versan sobre ese mundo imparable.

Debajo de la tierra hay incontables hormigas que marchan, conejos en sus madrigueras, gusanos retorciéndose e insectos siendo devorados. La computadora donde escribo atraviesa por diversos estados donde sus partes interactúan con el flujo eléctrico mientras el módem la provee de internet mediante una conexión inalámbrica. Inclusive, dentro de la inmóvil mesa se encuentran partículas subatómicas en acción, objetos infinitamente pequeños, similares a los millones de neutrinos que atraviesan nuestra piel cada segundo. Así, como constituyentes de este mismo dinamismo estamos nosotros, seres capaces de sentir el mundo, sus colores, sus texturas o alguna otra experiencia, ahí donde las palabras no bastan para evocarla. De este modo, sin duda afirmamos que somos *conscientes* pero ¿qué queremos decir con esto?

Por otro lado, cuando en diferentes campos de estudio se menciona la consciencia, usualmente ella viene acompañada por la afirmación fenomenológica: *consciencia es consciencia de algo*. Esto señala que la consciencia no es un evento ocurrido en el vacío. Es decir, no la concebimos sin tiempo, espacio o algún contenido. Vagamente entendemos que un ser consciente es aquel que tiene experiencias, y estas, nunca son huecas.

En consecuencia, parece que la condición de posibilidad de una experiencia incluye como mínimo, un experimentador y algo experimentado. Esto podría traducirse como un *ser que siente y algo que es sentido*, lo cual, nos inclinaría a postular una dualidad en toda experiencia.

Cuando veo la luna intuyo que el astro es algo que no soy yo. Si huelo un perfume distingo que la fragancia es algo ajeno a mí. Sin embargo, ¿en toda experiencia está presente esta diferencia? Cuando me pellizco siento dolor pero ¿la experiencia del dolor incluye ya al dolor mismo como algo mío?

Como objeción a esta dualidad podríamos suponer que, en cierto tipo de experiencias primitivas como el dolor, nunca identificamos dos cosas de arranque, es decir, al inicio no concebimos un sujeto y un dolor, sino que simplemente *sucede la experiencia del dolor*.

Posteriormente, cuando la memoria trae el testimonio de la experiencia dolorosa, esta la recordamos como un *yo* que sintió *dolor*. Así, parece que el *yo* emerge de un momento posterior a la experiencia misma, recuerdos que cuando se reproducen sustraen al *yo*, pero este nunca aparece explícito al momento de la sintiencia. ¿Acaso hay alguna experiencia donde el *yo* se manifieste desde el inicio?

¿Y si pensamos en la voz interna, el locutor que cada uno posee en su mente? ¿Eso no es ejemplo de una experiencia donde somos conscientes de nosotros mismos? En otras palabras, ¿acaso la voz de nuestras mentes no es un *yo* que se manifiesta desde el inicio? En semejanza con el dolor, la voz interna es una experiencia, algo que se siente.

En el momento de tener una *sensación*¹ no hay un *yo*, pues este es un concepto ulterior que surge de la diferencia ¿y dónde yace esta diferencia en el instante cuando siento? ¿Acaso este contraste entre *lo otro* y *yo* no es un momento posterior a la primera experiencia? Si esto fuera así, la voz interna sería la *experiencia de algo* —es decir, el contenido de una experiencia— que posterior a su manifestación reconozco como *yo* pero que no necesariamente tiene que ser así. Esto vendría apoyado por diversos testimonios, como el de los esquizofrénicos, quienes identifican como *otros* a esas voces que nacen de sus mentes, o bien, por los bebés de los que se sugiere no saben diferenciar dónde acaban ellos y empieza el mundo. Asimismo, este discurso sobre la *consciencia* en las *máquinas* depende, en gran parte, de aceptar que algunas experiencias no vienen siempre acompañadas de un *yo*.

Considerando este preámbulo ¿cómo definiría la consciencia un respetable diccionario de filosofía?:

Our own consciousness seems to be the most basic fact confronting us, yet it is almost impossible to say what consciousness is. Is mine like yours? Is ours like that of animals? Might machines come to have consciousness? Is it possible for there to be disembodied consciousness? Whatever complex biological and neural processes go on backstage, it is my consciousness that provides the theatre where my experiences and thoughts have their existence, where my desires are felt and where my intentions are formed. But then how am I to conceive the “I”, or self that is the spectator, or at any rate the owner of this

¹ De acuerdo con la definición de diccionario, las sensaciones son cosas como cosquillas, picazones, dolores, calor o frío sentidos, etc. La aparición de una sensación es la modificación más fundamental de la consciencia y comparte todos sus problemas relacionados con ella. Es usualmente pensada como un evento primitivo e interno, distinguido solo por el sentir crudo del poseedor. (Blackburn, 2005: 335)

theatre? [...] Even if we are convinced that consciousness somehow emerges from the complexity of brain functioning, we may still feel baffled about the way the emergence takes place, or why it takes place in just the way it does. (Blackburn, 2005: 74)

Como vemos, pensar la consciencia trae consigo una variada serie de problemas, sin embargo, la cuestión con la que parto es la última de esta cita. ¿Cómo y por qué se lleva a cabo la aparición de la consciencia?

El mencionado *Hard Problem of Consciousness* (Chalmers, 2014) refiere a una pregunta muy peculiar acerca de la consciencia. El punto de partida son las correlaciones que tenemos entre las experiencias humanas y aspectos físicos del cuerpo.

Por ejemplo, una explicación de la visión nos dice que en la retina se producen impulsos nerviosos, los cuales, a través del núcleo geniculado lateral del tálamo, llegan a la corteza visual del cerebro en el lóbulo occipital. La explicación puede convertirse más detallada involucrando términos específicos y explicaciones complejas de la neurociencia, sin embargo, teniendo en mente la sofisticación del sistema, reflexionemos: ¿Por qué todo este enramado neuronal viene acompañado de nuestra experiencia del mundo? “Why aren't we just robots who process all this input, produce all that output, without experiencing the inner movie at all?” (Chalmers, 2014)

De este modo, la solución al *HPC* busca una explicación que muestre satisfactoriamente la relación entre nuestras experiencias del mundo y las correlaciones que descubrimos con las distintas partes del cuerpo. La explicación tiene que ser una explicación causal donde la consciencia sea coherente con las demás entidades involucradas. En otras palabras, resolver el vacío explicativo requeriría de integrar *la consciencia* dentro de un conjunto de teorías aceptado. ¿Y cómo se logra eso? ¿Cuál sería la explicación definitiva que revele por qué somos conscientes?

Por otro lado, parece que un camino para entender la consciencia también requiere de estudiar las características con las que se presenta. ¿Qué contenidos tiene mi sintiencia? ¿Qué propiedades poseen mis experiencias?

Por las mañanas me levanto y lo primero que veo es el techo, percibo la suavidad de las cobijas y me siento algo somnoliento. Me percato de que hay espacio entre las cosas, que ellas se presentan con distintos colores y también soy consciente de algunos estados de mi cuerpo. La propiocepción me permite saber si tengo calentura o si algo anda mal con mi

estómago. Asimismo la consciencia parece ser esa representación del mundo que me acompaña todos los días. “Consciousness consists of all those states of feeling or sentience or awareness. It begins in the morning and it goes all day until you fall asleep or die or otherwise become unconscious.” (Searle, 2013)

Sin embargo, a pesar del rico contenido de mis experiencias, a pesar de mis múltiples capacidades como ser humano, la consciencia no es solo un cúmulo de representaciones aleatorias y desorganizadas.

Cuando le preguntamos a alguien que nos cuente sobre su día, normalmente nos platicará una pequeña historia donde luzcan los eventos más relevantes o de los que se acuerde. Por lo general estas narraciones son cortas y podemos agotarlas en pocos minutos. Pero, si intentásemos recordar las experiencias exactamente como las vivimos ¿cómo las contaríamos? ¿Detallaríamos cada escenario a pesar de que nunca observamos con atención los detalles? ¿Nos acordaríamos de todo lo que, en efecto, sentimos?

Ahora bien, ¿acaso nuestras experiencias —en el momento de tenerlas— son un enorme conjunto de datos definidos? En analogía con los reportes sobre nuestros días ¿no son las experiencias cotidianas también una especie de narración corta?

Es un hecho que en nuestras experiencias jamás nos enfocamos en todos los elementos del mundo, nuestra atención casi siempre está dirigida hacia cosas muy específicas.

Por ejemplo, si manejo en carretera mi cuerpo debería sentir el asiento, la textura del volante, la música, el olor del interior y quizá hasta mi respiración, además, mi visión abarcaría los retrovisores, la parrilla interior, el asiento del copiloto, el sol en el oriente y gran parte del cielo. A pesar de ello, todos estos elementos nunca están ahí, en realidad mi atención está casi solo en la carretera y el motor, todo lo demás, aunque aparentemente sigue ahí, nunca cae en el foco de mi consciencia² a menos de que ocurra algo que me obligue a cambiar el enfoque.

Another feature of consciousness is that it comes in unified conscious fields. So, I don't just have the sight of the people in front of me and the sound of my voice and the weight

² Podríamos pensar que estos aspectos en los que no prestamos atención de alguna forma son conscientes porque nuestro cuerpo continúa recopilando datos del exterior. Sin embargo, cuando afirmo que no caen en el foco de la consciencia quiero decir que en ningún momento ocurre una experiencia subjetiva donde se manifiesten estos datos.

of my shoes against the floor, but they occur to me as part of one single great conscious field that stretches forward and backward. (Searle, 2013)

“What distinguishes this preferred class of conscious representations from the much larger class of representations that never see the light of consciousness? (Churchland, 1997: 6)

Tomando esto en cuenta notamos que en las experiencias se discriminan los contenidos que se manifiestan. Pero no solo eso, a través de una reflexión sobre nuestros recuerdos podemos notar que estos contenidos no están acomodados a modo de una narración sin sentido; en cambio, las sensaciones que ahí surgen parecen contar pequeñas historias coherentes, narraciones unificadas que muestran una continuidad al representar nuevos hechos que devienen pasados. Coherentes porque en la mayoría de ocasiones no nos parecen extrañas las experiencias presentes en relación con las pasadas. Estas narraciones versan acerca de lo sentido, de nuestras experiencias cotidianas, de cómo los sucesos se nos presentan o de cómo representamos el mundo.

Hasta ahora vemos que la consciencia unificada, esta consciencia humana no se encuentra en cualquier lado, difícilmente ponemos en duda que esta sintiencia permanezca sin un cuerpo. Luego, sabemos que en el caso de los humanos, al cambiar ciertos aspectos de nuestra materialidad también ocurren cambios en nuestra consciencia, es decir, el tipo de sensaciones experimentadas cambia en relación con la configuración de nuestro cuerpo. Así, una cuestión esencial de la consciencia consistiría en explicar qué pasa cuando cambian las condiciones de nuestra percepción. Así, si modificamos aspectos de nuestro cuerpo y obtenemos *estados alterados de la consciencia*, es decir, estados que por su diferencia de contenidos cataloguemos como anormales, ya sea por reportes de las experiencias en primera persona o evidencias que tengamos para inferir que las experiencias cambian, ¿diríamos que seguimos siendo conscientes? ¿O es que, al experimentar contenidos clasificados como extraños, anormales o alterados, concluimos que ya no lo somos? Si un aspecto primordial de la consciencia es que se atraviesan por experiencias con contenidos, ¿estas suceden aun estando dormidos, exaltados o anestesiados?

But the actual fact turned out to be that consciousness is always associated with a functioning brain —not necessarily a normally functioning brain, since illusions of separability can be experienced during phenomena like the so called "out of body"

experience. During such states, however, the brain is still functioning, albeit in an altered state (Sullivan, 2006: 65).

Con esto último quiero mostrar que la forma de la consciencia no es siempre la misma. A veces es tan diferente como ver un atardecer en la playa y ver la puesta de sol durante un sueño, el contenido de ambas experiencias es rico, distinto, se siente diferente y, aún, ambas son manifestaciones de la consciencia.

As Kant (1787), Husserl (1913), and generations of phenomenologists have shown, the phenomenal structure of experience is richly intentional and involves not only sensory ideas and qualities but complex representations of time, space, cause, body, self, world and the organized structure of lived reality in all its conceptual and nonconceptual forms. (Van Gulick, 2021)

Si bien todos los días atravesamos por una serie variada de experiencias ¿por qué, en su mayoría, algunas nos parecen más anormales, más peculiares que otras? Por otro lado, nuestra variada serie de sensaciones revela que los contenidos de las experiencias son numerosos y, de alguna manera, podemos decir que cada experiencia es una forma de representación del mundo. A partir de esto, si asumimos que hay otros seres y estos experimentan contenidos distintos a los nuestros, ¿de qué depende esa diferencia en sus representaciones del mundo?

1.2 Otras formas de la consciencia

En el reino animal los marsupiales son mamíferos muy interesantes, ejemplos de ellos son los equidnas, los koalas y los canguros. Las hembras de estos últimos son capaces de amamantar a una cría al mismo tiempo que otro recién nacido ya se ha instalado en el marsupio, todo esto mientras un tercero se está gestando dentro de la madre. En contraste, los murciélagos poseen una habilidad llamada *ecolocalización*, la cual consiste en emitir ondas de sonido que rebotan en los objetos cercanos para devolver un eco que la criatura interpreta como información sobre el ambiente. ¿Sospechas cómo se siente una madre marsupial? ¿Crees que la imaginación nos haga sentir las cualidades de un canguro o un murciélago?

Si bien la experiencia de los humanos cambia al atravesar por *estados alterados de la consciencia* (alucinaciones, shocks, sueños, meditación, intoxicación por drogas, etc.) las diferentes representaciones del mundo no se agotan únicamente en la especie humana, hay un sinfín de seres que también poseen representaciones específicas y completamente diferentes entre ellos. “[O]ur normal walking consciousness, rational consciousness as we call it, is but one special type of consciousness, whilst all about it, parted from it by the filmiest of screens, there lie potential forms of consciousness entirely different.” (Shanon, 2003: 125)

De acuerdo con las diferentes condiciones a las que se adaptaron los antepasados de diversas especies, una consecuencia inevitable fue un cambio radical en la fisiología de todos ellos. Así, el cuerpo y las habilidades de un caimán se parecen poco a las de una abeja y, sin embargo, asumimos que ambos sienten.

Luego, parece que las diferentes formas de la consciencia vienen ligadas a un desarrollo evolutivo que permitió el desarrollo de distintas características. De este modo, la morfología de cada ser vivo está ligada con experiencias propias de cada estructura biológica, lo cual, permite diferentes representaciones del mundo.

Experiences seem no more able to exist without a self or subject to undergo them than could ocean waves exist without the sea through which they move. The Descriptive question thus requires some account of the self-perspectival aspect of experience and the self-like organization of conscious minds on which it depends, even if the relevant account treats the self in a relatively deflationary and virtual way. (Van Gullick, 2021)

En este sentido, la consciencia es la experiencia de cierta representación del mundo, “conscious experience is a type of virtual reality, a virtual world generated by our brain that constitutes a model of the real world.” (Westerhoff, 2015: 1)

En contraste con esto último debo remarcar que, aunque regularmente pensamos a los cerebros como los únicos órganos capaces de producir representaciones del mundo, hay seres vivos sin cerebro que también consideraríamos como sintientes, i.e. anémonas, medusas, erizos de mar, corales, entre otros.

El debate sobre cuáles criaturas lo son y cuáles no, es una cuestión en desarrollo y su resultado dependerá de los requisitos básicos que aceptemos para la atribución de consciencia, los cuales, no son claros. De cualquier manera, el hecho de observar en el mundo diferentes cuerpos y habilidades nos permite ver que el fenómeno de la consciencia se manifiesta en diferentes criaturas, en diferentes épocas, y no necesariamente estas distintas experiencias tienen las cualidades de la consciencia humana actual.

En vista de lo anterior y como consecuencia de la selección natural, a lo largo de las investigaciones se han encontrado cuerpos con características nunca antes vistas, características que pueden ser únicas de cierta especie, lo cual, si aceptamos a estas criaturas como sintientes, estas características les permitiría experimentar el mundo de una manera especial y desconocida para los humanos.

Como consecuencia, probablemente aun haya varios tipos de consciencia completamente desconocidos, formas de sentir de las que no tenemos idea. Quizá una mutación inesperada o una modificación deliberada den pie a nuevas formas de consciencia. Tal vez, ya sea voluntariamente o por accidente, surjan criaturas con una nueva perspectiva del mundo, aunque sus cuerpos sean metálicos o posean circuitos de silicio.

1.3 Ser ese organismo y no otro

En vista de lo anterior, ¿qué características posee un organismo para ser llamado consciente? Si bien como sinónimo de consciencia he utilizado las palabras *sintiencia* y *experiencia*, *un organismo consciente es aquel que simplemente siente o experimenta*. Como consecuencia, ya que por definición no existen las experiencias vacías o las sensaciones sin contenido, y estos contenidos los suponemos como representaciones del mundo, un organismo consciente es aquél que experimenta, al menos, una forma de representación del mundo.

But fundamentally an organism has conscious mental states if and only if there is something that it is like to be that organism- something it is like for the organism. We may call this the subjective character of experience. It is not captured by any of the familiar, recently devised reductive analyses of the mental, for all of them are logically compatible with its absence. It is not analyzable in terms of any explanatory system of functional states, or intentional states, since these could be ascribed to robots or automata that behaved like people though they experienced nothing. (Nagel, 1974: 436)

Aquí Nagel nos presenta al organismo consciente como alguien que experimenta lo que es ser ese organismo. En específico, hay algo que es ser ese y no otro. En analogía con esta exposición, la subjetividad manifestada revela al ser consciente, aquel que siente en primera persona, experimenta sus propias representaciones del mundo y no las del otro, pues en caso de no subjetivar sus experiencias, entonces, un sujeto podría sentir las sensaciones ajenas y esto nos llevaría a no saber dónde empieza un sujeto y acaba el otro.³

Sin embargo, confiamos en que esto no sucede y en que cuando creemos sentir las sensaciones de alguien más, en realidad solo nos imaginamos que ellos sienten como yo, y que si se comportan de cierto modo, seguramente sentirían algo como lo que nos imaginamos. Prueba de ello es que podríamos fabricar un robot con apariencia humana, especializado para aparentar que llora, después, algunas personas se imaginarían cómo se sentiría ese llanto aunque en efecto el robot no lo sintiera, o bien, lo que sintiera no tiene nada que ver con lo que imaginamos. Como corolario, ¿podemos decir que la consciencia siempre es individuada?

³En caso de ser así todo podría ser un mismo sujeto experimentando una misma consciencia, una teoría descabellada que podemos descartar para evitar la pérdida de la subjetividad.

There are occasions in which the sense of self dissipates but the experience of consciousness nevertheless remains. This suggests that consciousness as exhibited by individual selves may be a derivative phenomenon, which, in turn, depends on some sort of non-individuated consciousness. In other words, consciousness may perhaps precede individuated selves, the latter being formed when the former crystallizes into specific, bounded structures. (Shanon, 2003: 146)

Como Benny Shanon (2003) nos remarca, existen situaciones donde la sensación de un *yo* se pierde, aunque la consciencia —es decir, la experiencia o sintiencia de algo— se mantiene. En estudios realizados con Ayahuasca y otras sustancias como *mezcalina* o *LSD*, varios sujetos han reportado esta disipación del *yo*⁴, sin embargo, la disolución del *yo* también ha sucedido en prácticas como la meditación. “It has been repeatedly reported that one of the outcomes of sustained meditative practice and ecstatic states is the dilution of self-identity in a reality that encompasses and transcends the individual.” (Díaz, 2019)

Así, tanto el consumo de sustancias psicoactivas como otras prácticas mentales provocan alteraciones en la capacidad de autorreconocimiento, ya sea alterando los mecanismos cerebrales de identificación (suponiendo que existen), o alterando las sensaciones con las que nos identificamos, haciéndonos perder la usual noción del *yo*.

De cualquier modo, los estados alterados de la consciencia sugieren que esta, aunque por lo regular viene acompañada de un *yo*, no necesariamente tiene que ser individuada. No quiero decir que esta consciencia no individuada es una consciencia universal de corte *panpsiquista* que luego por algún mecanismo deviene individuada, sino que hay sujetos que al atravesar por ciertas experiencias no pueden identificar al sujeto que experimenta. Así, damos cuenta de sujetos cuyas experiencias no incluyen al *yo* que experimenta.

También, he de añadir que hay situaciones donde las experiencias son olvidadas, como en eventos traumáticos o sueños de los que tenemos certeza que existieron pero no recordamos su contenido. Por ello, tampoco es necesario que los agentes conscientes puedan recordar sus experiencias.

⁴ Aquí me refiero al “yo” como el mecanismo de autorreconocimiento. La capacidad de identificarme con alguna o varias representaciones mentales y no otras.

1.4 Consciencia emergente

Entre varias incógnitas, las preguntas *¿cuándo y cómo surge la consciencia?* han sido tema de debate por largo tiempo. Aquí se busca el momento donde podemos distinguir cuándo y cómo un objeto o una serie de objetos devienen conscientes en comparación con estados anteriores donde no lo eran.

Por ello, si aceptamos que en el pasado hubo seres que, aunque carecían de consciencia más tarde sus descendientes la desarrollaron, debería existir un momento donde las características de estos seres cambiaron y comenzó en ellos la consciencia.

Esto sugiere la existencia conjunta de ciertas características que, en el momento de presentarse —bajo la organización adecuada—, producen el fenómeno de la consciencia. “[C]onsciousness is just a property of matter at a certain stage of organization”. (Putnam, 1964: 680) “[T]hat persons are physical objects, and what makes an object a person is its functional organization.” (Pollock, 1990: 461)

Como consecuencia, muchos investigadores se han dado a la tarea de encontrar este conjunto de características, no solo comparando las funciones y estructuras de los seres conscientes con los supuestamente no-conscientes, sino también, intentando construir seres artificiales con las propiedades humanas. De este modo, dado que por su constitución los humanos son conscientes, si reunimos en un sistema suficientes características humanas, probablemente en algún momento emergerá una criatura consciente.

[B]y gradually increasing the sophistication of the software of the functionality of the robots in the series, we may be able eventually to arrive at genuinely conscious states. Everything neurophysiologists tell us about the causal bases of consciousness in humans and other animals suggests an extremely complex causal account quite at variance with that kind of picture. It is conceivable that the relevant neurophysiological factors will eventually be produced artificially. (Young, 1994: 236)

De acuerdo con este razonamiento, parece que en algunos lugares de la escala filogenética cesa la consciencia, pero ¿dónde? Tradicionalmente creemos que la consciencia es producida en el cerebro. Pensamos así puesto que, si perdemos un brazo, la lengua o los riñones, seguimos siendo conscientes. En cambio, si nuestro cerebro sufre un gran daño

perdemos la consciencia. Esto indica que las hipotéticas características suficientes para la consciencia deberían presentarse en el cerebro.

Dado este panorama, actualmente hay tanto neurocientíficos como filósofos quienes intentan encontrar cómo a partir de propiedades físicas determinadas puede emerger el fenómeno de la consciencia.

Such philosophers, by accepting the supervenience of o-experience [la consciencia de sujetos macroscópicos] on the physical, hope to accommodate o-experience in reality without adding to the physical facts. We can call such approaches physicalist solutions to the hard problem. (Goff, 2009: 291)

En 1998 el filósofo David Chalmers y el neurocientífico Christof Koch apostaron que en veinticinco años ya se habrían descubierto los correlatos neuronales de la consciencia, estos son, los constituyentes suficientes del cerebro que producen la consciencia. Sin embargo, en 2023 ambos se reunieron y el resultado fue incuestionable, se prorrogó la apuesta otros veinticinco años.

Hoy en día las actuales explicaciones acerca de *por qué somos conscientes* se han mostrado insatisfactorias, pues aún no han resuelto cómo surge la consciencia o por qué lo hace. No hay un consenso sobre las repuestas a pesar de haber estudiado la cuestión por décadas. Buscamos una explicación, quizá una que se ajuste a las palabras y a un número finito de cuartillas, pero ¿cómo vamos a explicar en párrafos algo tan fundamental como la consciencia?

En relación con la insatisfacción remanente de estas explicaciones parece muy razonable que busquemos las respuestas con otro modelo. Probablemente debamos suponer que la consciencia es una propiedad incluso más fundamental que el *espacio-tiempo*; quizá necesitemos un modelo que considere una interacción de distintos niveles de la realidad; o tal vez requiramos de postular la consciencia en todo, desde las partículas elementales hasta los seres más complejos. No obstante, a pesar de que filósofos como Donald Hoffman (2008), José Luis Díaz (2018), Tom Froese (2023) o Philip Goff (2009) han tomado estas rutas, el problema sobre *¿cómo surge la consciencia?* aún sigue abierto.

El problema sigue vigente porque, por ejemplo, incluso aceptando los presupuestos del *panpsiquismo estándar* o el *micro-panpsiquismo*, donde todos o algunos de los elementos últimos constitutivos de la realidad (e.g. fotones, quarks, electrones, etc.) instancien

propiedades fenoménicas, esto es, algún tipo de consciencia o protoconsciencia, surgen problemas explicativos ya que las experiencias de estas partículas tendrían cualidades diferentes a las experiencias de un organismo macroscópico consciente, lo cual, deja sin explicar cómo partículas elementales conscientes dan pie a sistemas más complejos con consciencia. Esto representa la cuestión ampliamente conocida como *The Combination Problem*.

En consideración de las últimas investigaciones en este campo, podríamos creer que estamos cerca de encontrar la solución a esta dificultad. Existen complejas y numerosas teorías que proponen diversas maneras para explicar distintos aspectos de la consciencia. Probablemente en los próximos años veamos avances significativos, no obstante, como veremos en los siguientes capítulos, el haberlos encontrado solo resolvería una parte de los problemas.

Capítulo 2

2.1 Representaciones mentales

“Does it make sense, in other words, to ask what my experiences are really like, as opposed to how they appear to me?” (Nagel, 1974: 448)

¿Cómo sabemos que somos conscientes? Tradicionalmente pensamos que las *representaciones mentales (RM)* —creencias, deseos, pensamientos o percepciones— son los paradigmas de la consciencia.

Los hechos como experimentar la imagen de mi cuarto o el olor del café son suficientes para convencernos de que existen tales representaciones, o al menos, una clase de experiencias que no se asemejan a nuestra denominación usual de las cosas. Las representaciones mentales contrastan con los objetos físicos, carecen de propiedades como tener color, dureza o volumen; pertenecen al conjunto de cosas que no podemos mostrar por ostensión, pues nadie puede señalar la creencia de que el rey de Francia es calvo. Si esto es así, entonces, ¿dónde se encuentran mis creencias?

Podríamos estar tentados a afirmar que nuestros pensamientos se encuentran en el cerebro, sin embargo, al abrir nuestros cráneos solo encontraríamos neuronas, tejidos y otros componentes biológicos. Si suponemos como Leibniz, en su *Monadología* de 1714, que tomamos un paseo por la cabeza y exploramos cada rincón, en ningún lado encontraremos *RM*. Claro que de esto no se concluye que las *RM* no están en los objetos materiales, sino que, aunque funcional o representacionalmente las podamos encontrar, el aspecto fenoménico de estas nunca se revela ahí.

A pesar de todo, damos cuenta de que existen estos objetos inmateriales, ellos son parte de nuestro sentido común puesto que los experimentamos a diario. No obstante, ellos encarnan uno de los grandes misterios de nuestra existencia. Si bien no podemos ubicar las representaciones mentales en el espacio parece que sí podemos situarlas en el tiempo. Observemos la siguiente tabla.

t_1	t_2	t_3
Un vehículo excede el límite de velocidad en el kilómetro diez.	La conductora nota que sus frenos no funcionan y entra en pánico.	El automóvil colisiona con la barrera de seguridad.

Los eventos t_1 y t_3 pueden ser ubicados tanto espacial como temporalmente puesto que el vehículo está en el kilómetro diez y asumimos que este se encuentra en una carretera. Asimismo, al colisionar no solo lo podemos ubicar en la vía sino también junto a la barrera de seguridad con la que se sufre el percance.

Sin embargo, el evento ocurrido en t_2 (el darse cuenta de una falla mecánica y entrar en pánico) no ocurre ni en el vehículo ni en el cuerpo de la conductora. No importa si agotamos la naturaleza física de la automovilista o del vehículo, no importa si revisamos átomo por átomo cada una de las cosas en este momento, el pánico —es decir, el pánico sentido—, no aparece ahí.

Dado que lo ocurrido en t_2 en efecto es el caso, nos vemos obligados a ubicar el terreno de su manifestación en un lugar diferente al de los objetos materiales. Podemos decir que todo esto ocurre en la consciencia, la cual, parece que no podemos ubicar espacialmente pero si temporalmente. Esto último solo tiene sentido cuando decimos que, o bien una experiencia sucedió antes, después o simultánea con algún otro suceso, como por ejemplo, el reloj del coche marcando las nueve.

Continuando con el ejemplo, si bien podemos predicar tanto del vehículo como de la conductora otras características como sus temperaturas, densidades y velocidades, estas son propiedades que no se pueden adscribir a los eventos mentales de t_2 .

Si hacemos el mismo análisis con cualquier otra *RM* (e.g., creencia de algo verdadero, el anhelo de paz, la sensación del dolor) veremos que estas no suceden en el espacio sino en la consciencia, en la subjetividad de cada agente.

¿Esto quiere decir que hay dos tipos de entidades? ¿No podemos abarcar ambas con una misma categoría ontológica? ¿El haber localizado las *RM* en la consciencia aporta algo para su explicación? “If there is a significant difference between an antirealist and realist theory of mental representation, this difference cannot merely consist in a difference between the ontologies of the two theories.” (Hellman, 1987: 4)

De acuerdo con los últimos párrafos, existen ciertas cosas de las que no podemos tanto ubicar espacialmente como adscribir otras propiedades físicas. Los sucesos mentales parecen tener una naturaleza tan diferente que en muchas ocasiones se les ha caracterizado con una ontología distinta.

En vista de ello, el dualismo surge como una explicación que toma en serio las diferencias para proponer una nueva categoría. Sin embargo, he de remarcar que la brecha entre lo físico y lo mental no es clara. “Dualist theories regard at least some aspects of consciousness as falling outside the realm of the physical, but specific forms of dualism differ in just which aspects those are.” (Van Gulick, 2021)

Recordemos que las visiones dualistas del mundo son antiguas y, tradicionalmente, los aspectos subjetivos han sido categorizados como parte del mundo espiritual. No obstante, lejos de hablar acerca de entidades místicas hablaré de lo *subjetivo* en tanto que refiere a *ser experimentado por individuos*.

Las experiencias que conciernen a estos individuos, también conocidas como *de acceso privilegiado*, poseen la cualidad de ser *privadas*. A través de palabras, imágenes, esquemas o lenguaje corporal, alguien puede intentar compartir sus representaciones mentales pero nunca vemos que alguien directamente las muestre. Así, las experiencias de dolor, sabores, colores o texturas son subjetivas y privadas en tanto que se comparten indirectamente, además de no ubicarse en el espacio físico. Asimismo, como remarqué en el primer capítulo, normalmente —más no siempre— estas son experimentadas por individuos/yos que se reconocen como tales.

Ahora, ya que conocemos algunas características en las que difieren los objetos físicos y mentales, luego ¿cómo explicamos su relación? Si bien aceptamos que existen objetos con ciertas propiedades, la explicación de estos no consiste solo en especificar qué propiedades poseen estando aislados, sino también, en explicar cómo interactúan con otros objetos.

Así, las explicaciones sobre lo mental no solo deben categorizar ontológicamente el contenido del mundo, también deben dar cuenta de cómo los estados mentales afectan otros sucesos —si es que lo hacen—.

We are left with three categories of mind-body theories which are relevant to the machine problem —Interactionism, Epiphenomenalism, and Materialism. Essentially, Interactionism holds that some mental events cause physical events, and some physical events cause mental events. Epiphenomenalism also admits the existence of mental events. It claims that no mental events cause physical events, no mental events cause

mental events, and all mental events are caused by physical events. [...] Materialism asserts that there are no mental events. (Thompson, 1965: 38-39)

Dennis Thompson nos señala que, a la hora de explicar las relaciones entre estados mentales y objetos físicos tenemos al menos tres teorías predominantes que pretenden cumplir esta tarea. Los siguientes puntos estarán dedicados al análisis de cada una.

2.2 Materialismo

En primer lugar consideremos al materialismo, cuyos adeptos sugieren que todo lo existente es materia. “By “materialism” I mean the theory that there is nothing in the world over and above those entities which are postulated by physics (or, of course, those entities which will be postulated by future and more adequate physical theories)”⁵. (Smart, 1963: 651)

De este modo, el materialismo se parece al fisicalismo en tanto que ambos se apoyan fuertemente en las teorías físicas como explicaciones últimas de la realidad. Con estos presupuestos parece que todo se podría explicar en términos de las entidades postuladas por esas teorías. “I mean by physicalism the thesis that a person, with all his psychological attributes, is nothing over and above his body, with all its physical attributes.” (Nagel, 1965: 339). He de resaltar que el materialismo difiere del fisicalismo en la medida en que este último posee una ontología más grande, pues el materialismo no parece tomar en cuenta de la misma manera que el fisicalismo la existencia de campos, fuerzas, estados cuánticos o la estructura espacio-tiempo; una postura rígida materialista consideraría que todo lo existente no es más que partículas materiales en movimiento.

Así, parece que el materialismo —entendido como que todo lo existente solo es materia— pretende abarcar todo bajo una misma categoría ontológica, lo material. Esto se postula principalmente para resolver el problema sobre cómo es que las representaciones mentales, entidades inmateriales —las cuales son radicalmente diferentes entre sí—, pueden interactuar con entidades materiales.

Usualmente se objeta que, si tenemos dos entidades cuyas propiedades difieren entre sí, no podrían interactuar ya que carecerían de la comunalidad necesaria para la interacción. Sin embargo, el materialismo, si bien no acepta la existencia de eventos mentales —o bien, estos son solo una manera de referirnos a objetos materiales de cierto tipo— evita el problema de la comunalidad al caracterizar ambos objetos como materia.

De cualquier modo, ya sea que los eventos mentales sean materiales o no, incluso los materialistas tienen que explicar por qué y cómo estas entidades interactúan con otros objetos.

⁵ En esta definición de Smart lo material no está compuesto solamente de partículas, sino también de otras entidades teóricas como la energía. Por ello, esto lo acercaría más a un materialismo moderno o a una especie de fisicalismo.

Esto es, ¿cómo mi sensación de hambre afecta o no a los demás objetos del mundo para que más tarde mi cuerpo se provea de alimento?

Por otro lado, si imaginamos que en la *tierra gemela*⁶ existe alguien materialmente idéntico a mí pero carece de representaciones mentales, entonces, esto quiere decir que, o bien las características materiales no son suficientes para determinar los estados mentales —y por tanto el materialismo es falso—, o estas características materiales son indeterministas, por lo que a veces pueden producir o no, un estado mental. En este sentido, si existiesen estos individuos materialmente idénticos, comúnmente llamados *zombies filosóficos*, dado que tendríamos dos sistemas físicamente idénticos cuyos outputs son distintos, inferimos que incluso aún teniendo una descripción completa de los estados materiales esto no es suficiente para describir las *RM* ligadas a ellos. O bien, si suponemos que estos sistemas son genuinamente aleatorios —y por tanto impredecibles— entonces puede ocurrir que ambos sistemas, a pesar de ser idénticos, difieran en un momento y el desarrollo posterior de sus cadenas causales sea distinto.

Luego, a pesar de tener dos sistemas microfísicamente idénticos puede ocurrir la casualidad de que, en casi todo el desarrollo temporal de los dos mundos, ambos eran idénticos, no obstante, en algún momento alguno produjo consecuencias distintas al otro, específicamente en lo concerniente a la aparición de *RM*. No obstante, esto último refiere a posturas que el materialista o el fisicalista difícilmente aceptarían.

For the physicalist, whatever her stripe, must at least believe that the microphysical facts determine all the facts, that any world that was exactly like ours in all microphysical respects (down to the smallest detail, to the position of every single boson, for example) would have to be like our world in all respects (having identical mountains, lakes, glaciers, trees, rocks, sentient creatures, cities, and so on). (Tye, 2021)

Entonces, si bien el materialismo desea comprimir todo bajo la categoría de lo material, frente al planteamiento de una tierra gemela sin *RM*, el materialista respondería que no puede existir una tierra gemela con tales propiedades ya que lo material determina todo, y si tenemos dos mundos con los mismos objetos materiales, como consecuencia no podríamos tener sistemas diferentes.

⁶ Un planeta hipotético materialmente idéntico al nuestro.

Por otro lado, si el planteamiento se acepta cómo posible, se podría salvar el materialismo mostrando que el mundo es indeterminista, y que, si bien tenemos dos mundos materialmente idénticos, las propiedades mentales de uno podrían estar ausentes ya que no es necesario que dos sistemas físicos arrojen siempre los mismos resultados; debido a su indeterminación los resultados podrían variar.

Con respecto a estas respuestas, la primera parece sólida puesto que, en efecto, el mundo en su mayoría parece comportarse de manera determinista. Generalmente, cuando tenemos dos sistemas idénticos la variación en los resultados es nula o poca.⁷

En cuanto a la segunda, parece bastante forzada debido a que postula dos sistemas que, si bien se han comportado del mismo modo por un extenso tiempo, la aleatoriedad se presenta en un momento tardío y, además, esta se manifiesta de modo que mantiene las propiedades del sistema con excepción de las representaciones mentales.

Además, al suponer que tenemos dos mundos que, a pesar de tener *RM* en uno, ambos se mantienen físicamente idénticos, esto supondría que las *RM* no provocan cambio alguno en los sistemas materiales, lo cual, nos acercaría hacia una postura *epifenomenalista*.

⁷ Se sugiere que el mundo se comporta de manera determinista en vista de que, al analizar los sistemas de aleatoriedad encontramos impredecibilidad en vez de aleatoriedad genuina. Por ejemplo, en el lanzamiento de una moneda podemos conocer el estado final si conocemos todas las variables involucradas (i.e. fuerza inicial, peso de la moneda, resistencia del aire, velocidad de rotación, etc.) (Diaconis, 2007) Asimismo, en la mecánica cuántica la ecuación de onda se trata de manera determinista y el hecho de no poder determinar simultáneamente ciertos pares de magnitudes de una partícula se debe a una barrera epistémica (impredecibilidad) mas no a aleatoriedad genuina. Por ello ha sido posible postular teorías cuánticas de variables ocultas, sin embargo, estas no han sido muy exploradas debido a que para fines prácticos superar el principio de indeterminación es irrelevante. Ahora, hay sistemas caóticos, los cuales con el tiempo presentan cambios drásticos a pesar de tener condiciones iniciales muy similares. No obstante, esto no aporta nada en contra del determinismo ya que, si tuviésemos dos sistemas caóticos con las mismas condiciones iniciales (lo cual es prácticamente imposible), su desarrollo sería idéntico. (Rojas, 2020)

2.3 Epifenomenalismo

“Content passes through one’s mind, but one does not experience oneself as generating it.” (Shanon, 2003: 133)

El epifenomenalismo es la postura dualista que admite la existencia de entidades mentales pero sugiere que ellas no tienen ningún rol más allá de ser experimentadas. Bajo este marco, las *RM* serían únicamente efectos, *epifenómenos*, casos terminales de cadenas causales puesto que no pueden ser causas de otros fenómenos.

Parece algo contraintuitivo en vista de que cotidianamente explicamos sucesos involucrando elementos mentales como *causas* y cambios en el mundo como *consecuencias*. Por ejemplo, decimos que la sensación de frío nos mueve para cobijarnos, quemarnos el dedo nos aleja del fuego, escuchar una canción puede hacer que bailemos y sentir el motor de nuestro auto causa que subamos la marcha.

En vista de ello, ¿por qué tiene apoyo esta postura si, *a prima facie*, se muestra contraintuitiva? Gran parte de este soporte viene dado por *los experimentos de Benjamin Libet*, los cuales, parecen poner en duda la función de la consciencia como elemento causante.

El experimento más controversial se desarrolló a finales del siglo pasado (Libet, 1985) y trató de investigar acerca del momento exacto en que las personas toman una decisión consciente en relación con la actividad cerebral que les precede.

Para este experimento se pedía a los participantes mover su muñeca en el momento que ellos lo desearan, siempre y cuando prestaran atención al instante exacto donde decidían hacerlo. Esto lo medían a través de un reloj que ponían frente a los participantes.

De este modo hubo tres elementos que se tomaron en cuenta, el primero, el *potencial de preparación (RP)* que, a través de un electroencefalograma se medía la actividad cerebral previa al movimiento. Segundo, se registraba el momento en que la muñeca del participante se movía. Y tercero, se tomó en cuenta el momento en que los participantes eran *conscientes* de su decisión sobre mover su muñeca.

The initiation of the freely voluntary act appears to begin in the brain unconsciously, well before the person consciously knows he wants to act! Is there, then, any role for conscious will in the performance of a voluntary act? (see Libet 1985): To answer this it must be recognized that conscious will (W), does not appear about 150 msec. before the muscle is activated, even though it follows the onset of the RP. An Interval of 150 msec

would allow enough time in which the conscious function might affect the final outcome of the volitional process. (Actually, only 100 msec. is available for any such effect. The final 50 msec. before the muscle is activated is the time for the primary motor cortex to activate the spinal motor nerve cells. During this time the act goes to completion with no possibility of stopping it by the rest of the cerebral cortex.) (Libert 1999, p. 51)

Actualmente los resultados de esta investigación son debatidos y ponen en duda el rol tanto de la consciencia como del libre albedrío. Se argumenta que, a pesar de tener presente el fenómeno de la consciencia, esta ocurre demasiado tarde para ser considerada la causa de las acciones en vez de su resultado —o a lo mucho un efecto conjunto de una causa inconsciente que le precede—.

Así, la consciencia parece ser el efecto de las operaciones cerebrales, “[...] like a post facto printout or the result displayed on one's computer screen than like the actual processor operations that produce both the computer's response and its display.” (Van Gulick, 2021)

Bajo este marco, parece que se realiza un complejo proceso biológico, principalmente en el cerebro, y más allá de producir los outputs correspondientes, cierta parte de este proceso viene acompañada de representaciones mentales. ¿Por qué? ¿Qué liga estos disparos neuronales, estos tejidos y material biológico con la película de lo mental?

“... [W]hile constituting only 2% of body weight [the brain], draws more than 20% of the organism's total energy (Attwell & Laughlin, 2001)”. Si seguimos los presupuestos del epifenomenalismo, diríamos que el cuerpo lleva a cabo un complejo proceso que utiliza una gran parte de la energía total del cuerpo para producir representaciones mentales que no afectan nada en las cadenas causales posteriores de eventos. Es decir, el epifenomenalismo sugiere que, dado que las representaciones mentales no son causas de algo, entonces no existirían diferencias si tuviésemos una tierra gemela sin *RM*.

Por lo tanto, independientemente si el epifenomenalista acepta como posible la suposición de esta tierra gemela, en caso de no haber *RM* todos los sistemas físicos se mantendrían exactamente igual; lo único que se perdería serían las películas subjetivas. No obstante, esto parece contraintuitivo porque de este modo todas las acciones humanas se llevarían a cabo del mismo modo incluso sin la consciencia de los actos realizados.

Por otro lado, también perderíamos explicaciones evolutivas del por qué tenemos experiencias conscientes. Si las *RM* no son causas, entonces, no habría ninguna ventaja

evolutiva para ninguno de los organismos que hayan desarrollado consciencia. Esto quiere decir que los cuerpos de los organismos conscientes realizarían procesos de alto costo energético para producir experiencias, sensaciones, las cuales no tendrían utilidad alguna.

Si aceptamos las implicaciones de la teoría epifenomenalista, tendríamos eventos mentales estériles, explicaciones de sentido común falsas (como afirmar que mi sensación de enojo provocó un grito), y una teoría que no puede dar cuenta de los beneficios evolutivos de poseer cierta propiedad (la consciencia) emparentada con procesos biológicos. Así, sería aún más misterioso por qué tenemos consciencia en vez de no tenerla y el epifenomenalismo aportaría poco o nada para resolver el *Hard Problem of Consciousness*.

Finalmente, si las máquinas serán o ya son conscientes, de acuerdo con el epifenomenalismo no podríamos detectar si hay o no consciencia a través de los efectos de sus representaciones mentales en los sistemas físicos, lo cual, cerraría una vía experimental para dar cuenta de otras mentes.

Los experimentos de Benjamin Libet aún son materia de discusión ya que su interpretación no es definitiva. No hay consenso en que ellos demuestren el éxito del epifenomenalismo. Incluso Libet menciona que la consciencia del acto a realizar conlleva una capacidad de veto, una capacidad para detener una acción que parecía haber empezado cuando se detectó el *potencial de preparación*. Esto también sugeriría que la consciencia, cuando se presenta, puede cambiar el estado final de un evento, lo cual nos alejaría del epifenomenalismo clásico y nos acercaría a una postura interaccionista.

2.4 Interaccionsimo

De acuerdo con un estudio sobre la precisión introspectiva (Fleming, 2010), las personas que mostraron mayor habilidad para pensar sobre sus propias acciones tenían una porción más grande de córtex prefrontal, el cuál fue detectado por escaneos cerebrales. Sin embargo, ¿esta porción mayor se debe a que las personas eran más capaces para pensar sobre sus acciones o el hecho de tener un córtex prefrontal más grande provocó el desarrollo de esta habilidad?

El *Interaccionismo* se caracteriza como una postura dualista que acepta la existencia tanto de entidades materiales como mentales y, además, sugiere que ambas interactúan. En este sentido las *RM* producirían cambios específicos en los sistemas físicos, así como estos mismos sistemas dan pie a los eventos mentales.

Aunque actualmente se considera que hay pocos investigadores bajo la etiqueta de *interaccionistas*, esta propuesta es relevante puesto que deja abierta una vía experimental para detectar los cambios que sufriría un sistema dado que este posee consciencia.

Por otro lado, si esta teoría propone que los resultados finales de ciertos eventos no son independientes de las *RM* involucradas, esto vendría de la mano con el sentido común. “Intentions and decisions involve selection mechanisms among several alternatives that become necessary to plan and direct programs of action deemed efficient or valuable.” (Díaz, 2018: 214)

De este modo, procesos como el de deliberar acerca de una decisión no son gratuitos, pues, debido a que estos involucran una serie de representaciones mentales que modelan estados futuros posibles del mundo, y estas representaciones provocan un cambio en el resultado final de la deliberación, las *RM* ahora formarían parte de la cadena causal de eventos en tanto causas de estados subsecuentes.

Esto contrasta con el epifenomenalismo en tanto que las *RM* ya no son estériles, por lo cual, una explicación evolutiva de la consciencia sería posible.

Si bien en el primer capítulo mencioné que el tipo de *RM* depende de la morfología del organismo, entonces, los cambios graduales tanto en el cerebro como en el resto del cuerpo también habrían modificado la vida mental de los agentes en tanto que esta devino más adecuada para la supervivencia de la especie.

Esto sugiere que tendríamos representaciones mentales primitivas en los ancestros más lejanos, las cuales, cambiaron gradualmente hasta llegar a los aparatos cognitivos de los

organismos actuales. Las criaturas cuyos procesos mentales servían para ponderar la reproducción de su especie, tuvieron una ventaja durante su desarrollo evolutivo.

Como ejemplo pensemos en la memoria, una serie de mecanismos cerebrales enraizados en diferentes componentes biológicos, cuyos contenidos útiles y significativos no son accesibles sino mediante la consciencia, una serie de representaciones o metarepresentaciones mentales sobre el pasado que son relevantes para la actualidad del agente. “Memory is a fundamental cognitive property of all animals and is essential for adaptive behavior. It allows past experience to modulate present behavior in an uncertain environment, thereby increasing an organism’s likelihood of survival and fitness.” (Ryan, 2020: 207)

Los contenidos mentales que son de ayuda para el organismo solo serían relevantes para cambiar el estado final de un evento si, mediante su representación consciente, el organismo tiene acceso a ellos y, además, utiliza su información en concordancia con las condiciones ambientales.

Por otro lado, si bien el interaccionismo encaja con una explicación evolutiva aún existen dificultades. El *problema de interacción causal* propone el desafío de aclarar cómo es que una substancia inmaterial provoca alteraciones en sistemas materiales. Creo que más allá de postular una misma ontología (monismo) para obtener la comunalidad necesaria para la interacción, lo único que puede dar cuenta de tal interacción es la correlación causal entre eventos.

A propósito, este mismo problema le fue presentado a Descartes en su correspondencia con Elisabeth de Bohemia, a lo cual, entre varias cosas, él responde que la interacción entre *mente* y *cuerpo* es similar a la interacción entre la *gravedad* y los *objetos*. Si bien la gravedad no es algo material, sí tiene repercusiones en los objetos físicos.

¿Qué importa si les ponemos la etiqueta de material o inmaterial? El hecho es que vemos una correlación entre ellos, cuando uno se presenta, siempre —o casi siempre— ocurren cambios en el otro. Así, podemos tener explicaciones que involucren el *espacio-tiempo*, una entidad inmaterial en la forma de un marco geométrico que, dependiendo sus condiciones, provoca alteraciones en objetos como la luz.

Por un lado, el interaccionismo propone una relación entre entidades mentales⁸y materiales pero no nos dice específicamente cómo ocurre esto. Es decir, ¿hay leyes para explicar la interacción entre *RM* y distintos sistemas físicos? Algunas propuestas sugieren que las *RM* tienen el poder causal de cambiar los procesos físicos mediante una transferencia de energía, sin embargo esto conlleva nuevos problemas. “For example, if causal power was flowing in and out of the physical system, energy would not be conserved, and the conservation of energy is a fundamental scientific law.” (Robinson, 2023)

De igual modo, se han propuesto diferentes soluciones para evitar el conflicto con la *ley de la conservación de energía*, como sugerir que las *RM* no cambian la cantidad de energía del sistema sino su distribución, o suponer que esta ley no se viola porque el cuerpo humano no es un sistema cerrado en vista de su interacción con el ambiente. Incluso se sugiere que hay que revisar el concepto mismo de energía.

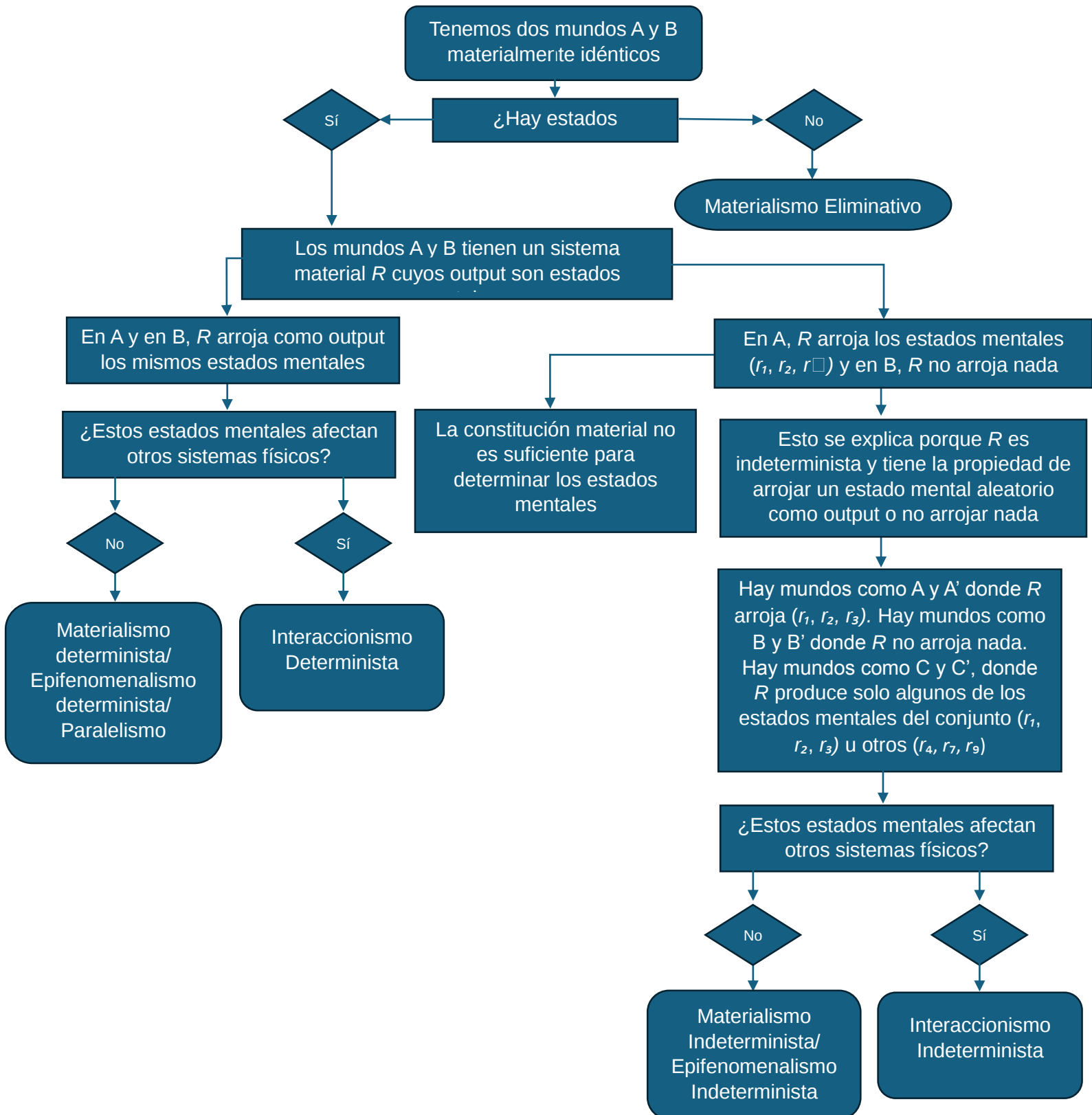
En general, el interaccionismo parece ofrecer diferentes beneficios teóricos en comparación con el materialismo o el epifenomenalismo. Esta teoría se acopla bien con las explicaciones de sentido común, es coherente con un esquema evolutivo, propone vías para evitar el problema de interacción casual y, sobre todo, deja abierto el campo experimental para investigar sobre cómo, bajo que principios y en qué medida, las *RM* modifican los sistemas físicos.

Si bien, debido al carácter subjetivo de las *RM* es difícil relacionar con precisión las entidades mentales y las materiales, esto no es un impedimento para continuar nuestra investigación. “You can have a completely objective science, a science where you make objectively true claims, about a domain whose existence is subjective.” (Searle, 2023)

Si en el futuro la teoría resulta exitosa y nos permite dar mejor cuenta de la relación entre *RM* y diversos sistemas físicos, se vislumbra la posibilidad experimental de acercarnos para conocer, si las máquinas poseen o no, una vida mental.

⁸ Ya sea que consideremos a estas como sustancias mentales, como propiedades o modos de un sujeto.

Ilustración 1. Tipos de dualismo



2.5 The Hard Problem of Consciousness

The centerpiece of the science of consciousness in recent years has been the search for correlations, correlations between certain areas of the brain and certain states of consciousness. [...] Now we understand much better, for example, the kinds of brain areas that go along with the conscious experience of seeing faces or of feeling pain or of feeling happy. But this is still a science of correlations. It's not a science of explanations. (Chalmers, 2014)

En las investigaciones acerca de la consciencia siempre hay un tema que incomoda, contrasta, sonsaca, nos hace dudar, nos hace pensar si esta propuesta plantea una cuestión genuina o si solo es una ilusión del lenguaje; el *Hard Problem of Consciousness (HPC)* bosqueja una pregunta como: ¿Por qué todos estos procesos, del cerebro, de mis brazos, de mi cuerpo, vienen ligados con la experiencia, con esta subjetividad?

La manera de etiquetar este misterio fue acuñada por David Chalmers pero, actualmente, su empleo va más allá de su investigación. El vacío explicativo que hay entre los elementos físicos y la sintiencia, entre los componentes biológicos y la vida mental, ha sido tema de interés a lo largo de años. Sin embargo, tampoco estamos cerca de una respuesta definitiva. A la incógnita planteada en el párrafo anterior, ¿cuál sería la respuesta adecuada?

La interrogante se enunciaría como: ¿Por qué hay experiencias en vez de no haberlas? A lo que podríamos contestar que, desde hace mucho, hemos atribuido la capacidad de sentir a los seres que tienen ciertas características, como tener un sistema nervioso o pertenecer al reino animal y, si no hubiese tales animales entonces no habría sintiencia. Así, la pregunta podría transformarse en: ¿por qué existen los animales sintientes en vez de no haberlos? Entonces, la respuesta quizá requiera de buscar las causas necesarias que provocaron la existencia de los animales. Así, ¿dónde detenemos la explicación causal?

No, ya nos alejamos del propósito inicial, por consiguiente, reformulemos, ¿por qué ciertas experiencias vienen ligadas con ciertos mecanismos específicos del cerebro? Quizá podamos responder que cierta experiencia viene ligada a cierta zona de actividad cerebral porque no hemos visto lo contrario, es decir, hemos observado que la visión aparece emparentada con los ojos, fotorreceptores, nervios ópticos y señales eléctricas en el córtex visual y no que la visión está vinculada con los músculos del abdomen. O sea que, lo que liga ciertas experiencias con ciertas zonas de actividad cerebral es que así lo hemos visto, la

continuidad; cada vez que sucede una, también sucede la otra. Lo que une estos eventos es que parece haber un vínculo causal entre ellos apoyado por repeticiones observadas a lo largo del tiempo.

Sin duda esta respuesta parece insatisfactoria porque nos remonta al problema de la inducción. ¿Solo porque dos eventos ocurren seguidos en una cadena causal, implica que así seguirán siendo o que hay una ley invisible que hace necesario este vínculo? Si hasta ahora solo hemos encontrado diamantes con un peso menor a 700 g, ¿esto sugiere que no pueden existir diamantes más pesados? ¿Hay una ley que limite el peso de los diamantes encontrados?

Por otro lado, cuando intentamos descifrar qué es el dolor, las explicaciones no se limitan a la conducta visible o a dar una simple correlación entre un lugar específico del cerebro y un número contable de neuronas.

[A]lthough the meaning of 'pain' may be explained by reference to overt behavior, what we mean by 'pain' is not the presence of a cluster of responses, but rather the presence of an event or condition that normally causes those responses. (Of course the pain is not the whole cause of the pain behavior, but only a suitably invariant part of that cause, but, similarly, the virus-caused tissue damage is not the whole cause of the individual symptoms of polio in some individual case, but a suitably invariant part of the cause.)
(Putnam, 1963: 27)

De este modo, me parece que una explicación completa del dolor tiene que considerar todos los elementos invariantes que puedan agregarse exitosamente en una cadena causal explicativa. Por ejemplo, para el dolor debemos integrar las diferentes partes del cuerpo involucradas (no solo el cerebro) y su relación causal en torno a esta experiencia. Al mismo tiempo, también deben mencionarse las características del dolor sentido, no solo desde una objetividad en tercera persona, sino también desde la sintiencia en primera persona, la consciencia de los sujetos. A su vez, si creemos que las virtudes epistémicas tradicionales — tales como la simplicidad, la consistencia o la fertilidad — son una guía importante para ampliar nuestro alcance epistémico, es menester integrarlas en nuestra explicación. Incluso para expandir nuestra explicación podríamos suponer situaciones contrafácticas donde hay dolor pero no conducta del dolor, o donde se presentan los elementos físicos del dolor pero no la experiencia de este.

¿En qué se diferenciaría esta explicación en contraste con la respuesta esperada al *HPC*? ¿Hay algo que dejamos sin aclarar? ¿Existe algo como la interpretación definitiva de la consciencia?

Recordemos que, en el capítulo uno, mencioné una característica esencial de la consciencia, la cual dice que consciencia siempre es consciencia de algo, tiene un contenido, luego, esto refiere siempre a experiencia, sintiencia o consciencia de algo. En otras palabras, por consciencia nos podemos referir a la capacidad para sentir, experimentar o ser conscientes de algún contenido, sin embargo, esta capacidad no existe en abstracto, no existe la consciencia absoluta sino como una categoría lingüística para referirnos a la capacidad de atravesar por toda una gama de experiencias.

No obstante, tal capacidad siempre se manifiesta como una experiencia determinada, una sensación que es esa y no otra, por lo cual, no es posible dar una explicación de la consciencia en abstracto. Dado que la consciencia siempre se manifiesta de cierta forma y con cierto contenido, las explicaciones tienen que versar sobre todo lo concerniente a tal contenido.

En cambio, si el *HPC* intenta descifrar por qué existen correlaciones entre todo el sistema físico que acompaña una experiencia, entonces me parece que refiere a una pregunta metafísica sin respuesta factible. Una pregunta similar a ¿por qué hay algo en vez de nada?

Brute regularities exist in the other sciences. At a certain point we simply have to accept that the physical world works in a way that we can describe but cannot explain. For example, the behavior of superstrings or elementary wave-particles can be described but not explained—it is the starting point for higher level physical explanations. Scientists cannot say why physical brute regularities exist. They are simply how the universe works. (Gamez, 2018: 41)

Si bien el *HPC* nos incita a criticar la mera correlación de eventos, al mismo tiempo nos hace cuestionarnos ¿qué más queda por explicar? En consideración de la historia de la filosofía, dada la poca convención y fertilidad que podríamos esperar al intentar responder cuestiones con este carácter metafísico, no creo que el propósito de resolver el *HPC* sea el descubrir en su aspecto metafísico por qué hay correlaciones en vez de no haberlas, sino cuáles son las características de estas correlaciones, su matematización, sistematización, y todo lo que podamos decir sobre nuestra interacción con ellas.

And I have only one real message in this lecture, and that is, consciousness is a biological phenomenon like photosynthesis, digestion, mitosis —you know all the biological phenomena— and once you accept that, most, though not all, of the hard problems about consciousness simply evaporate. (Searle, 2013)

2.6 Reduccionismo

There are some subjects, economics and politics included, in which too simplistic a point of view is equivalent to a catastrophe. Unfortunately, one is faced with the same dilemma in trying to describe a system consisting of 10 followed by 10 zeros' worth of neurons, 10 followed by 15 zeros' worth of connections, and innumerable other details, some of them poorly understood. (Eldeman, 1982: 5)

Cuando investigamos sobre la consciencia nuestro objetivo es producir explicaciones satisfactorias que dejen aclarados todos los misterios, preguntas y objeciones, sin embargo, estas explicaciones no son uniformes y tampoco se apegan a la misma estructura en todos los discursos. La jerga, los tecnicismos, la cantidad, distribución y tipo de información, la clase de preguntas, todos estos son elementos que varían a pesar de estar dirigidos hacia el mismo tema.

En vista de este rango de variables, de las características heterogéneas y de los diferentes modos de sumergirnos en los problemas de la consciencia, consideremos que las diversas disciplinas que la estudian establecen los parámetros que consideran adecuados, sensatos, acorde con sus fines, y estos, guían sus investigaciones.

La filosofía, por ejemplo, normalmente sigue una estructura lógico-argumentativa, y es a través de herramientas como esta que realiza análisis en diversas áreas. Por otro lado, las ciencias también tienen sus propios parámetros, herramientas, métodos, tecnicismos, modos de presentación y maneras de experimentar, las cuales difieren gradualmente de una ciencia a otra. Así, podemos pensar que la física es más similar a la química que a la biología, o que la filosofía es más cercana a la psicología que a la geología. Asimismo, dentro de cada disciplina a la vez tenemos subdisciplinas que también difieren entre sí, como la filosofía de la neurociencia y la fenomenología o la biología molecular y la ecología.

Independientemente de sus semejanzas, cada disciplina se desarrolla con las características propias de sus prácticas. Sin embargo, existen ciertas tendencias que pretenden explicar todo bajo una misma disciplina, como si los diferentes fenómenos de la realidad se redujesen a un solo campo de estudio.

Este paso reductivo tiene mucho sentido, pues diversos sistemas solo se entienden mediante el estudio detallado de sus partes constitutivas. Por ejemplo, el funcionamiento de un organismo al nivel celular —el cual principalmente se basa en comprender la secuencia

de *ADN*, los genes específicos y los aminoácidos involucrados en la síntesis de proteínas—nos permite predecir rasgos genéticos, mutaciones y cambios básicos que llegan a afectar rasgos fenotípicos y funcionales. Del mismo modo, los computadores funcionan mediante la combinación de numerosas operaciones diminutas ejecutadas por transistores. Asimismo están involucrados diversos componentes como las memorias, monitores, procesadores, circuitos o tarjetas madre, partes esenciales para entender el funcionamiento de un ordenador.

Así, a partir de una visión reduccionista se puede llegar a la idea de que cualquier fenómeno puede ser explicado mediante la disciplina que estudie las partes minúsculas constitutivas de un fenómeno. De este modo, la física de partículas y la física cuántica son las ciencias ideales para un reduccionista.

Sin embargo, mientras que una visión reduccionista nos puede encaminar al mejor entendimiento de un sistema, también nos puede llevar a evadir puntos de vista que, aunque no se enfoquen en los componentes mínimos de este, son más valiosos para una explicación adecuada. “Physics presumably has as little use for the categories of macrobiology as it does for the categories of commonsense psychology; it dissolves the behavior as well as the behavior. What's left is atoms in the void.” (Fodor, 1987: 9)

Por ejemplo, durante largo tiempo los trastornos mentales como la depresión fueron atribuidos únicamente a desbalances químicos en el cerebro, lo cual llevó a crear tratamientos basados fuertemente en medicación sin tomar en cuenta los aspectos psicológicos y sociales. Así, los pacientes combatían este desbalance pero no modificaban su estilo de vida, relaciones traumáticas u otros aspectos cognitivos que también formaban parte de la depresión.

Por otro lado, en la conservación de ecosistemas han ocurrido casos donde el propósito es proteger a una especie en peligro sin considerar el ecosistema completo que habita. Esto ha llevado, por ejemplo, a salvar depredadores sin asegurar la supervivencia de sus presas, lo cual deriva en un desbalance. Una conservación eficaz requiere de comprender las relaciones interdependientes dentro del ecosistema en lugar de centrarse en las especies aisladas.

Instead of taking things apart in order to understand the separate components, systems thinking focuses on process and performance on how the parts interact together to make up a functioning system. According to the systems theory, a system is an indivisible

whole where the performance of the whole entity is affected by every one of its parts and the way that any part affects the whole depends on what at least one other part is doing. (McLellan, 1988: 3)

Debido a estos casos donde la reducción no resulta favorable, podemos dar cuenta de propiedades en un sistema que solo aparecen desde cierto punto de vista, propiedades que no se encuentran en los componentes más pequeños o básicos, sino que parecen irreducibles.

In a sense, the seeds of this objection to the reducibility of experience are already detectable in successful cases of reduction; for in discovering sound to be, in reality, a wave phenomenon in air or other media, we leave behind one viewpoint to take up another, and the auditory, human or animal viewpoint that we leave behind remains unreduced. (Nagel, 1974: 445)

Esta misma tendencia a reducir aspectos de la realidad nos hace notar que hay elementos irreducibles, propios de cierta perspectiva o punto de vista que está ligada con las propiedades detectables de un sistema. Ahora, en el caso de la consciencia, cuando hablamos de sensaciones como la euforia o la ansiedad, no podemos reducirlas a términos puramente químicos. Existe la experiencia del sujeto, privada, que a pesar de ello no elimina el vínculo con las propiedades materiales del cuerpo.

Aquí resalta un punto relevante que comúnmente se pasa por alto: la intraducibilidad de las experiencias en términos de cierta disciplina. Dentro de algunas explicaciones, las cualidades de estas experiencias son nombradas como *qualia*, que me limitaré a referirme así a ellas en tanto tipos de sensaciones. Luego, si nuestro propósito es estudiar la consciencia y por consiguiente los contenidos emparentados a ella en forma de *qualia*, ¿cómo vamos a develar su naturaleza sino mediante la traducción de estos a términos inteligibles?

Precisamente por su característica individual, privada, subjetiva, no existe la publicidad en el estudio directo de los *qualia*, sin embargo, esto no impide nuestro acercamiento, sino que lo limita a la interacción mediante su referencia en términos de un lenguaje. El problema con el reduccionismo es que pretende abarcar tal acercamiento mediante un único tipo de perspectiva, dejando afuera otras representaciones que puedan tener un valor epistémico.

Si tomamos en consideración esta postura más amplia, aceptando que la mejor representación lingüística de un suceso mental puede provenir de distintos campos de estudio, entonces la mejor terminología dependerá del tipo de práctica que realicemos.

En otras palabras, si queremos compartir cómo se siente saborear un té verde —que por supuesto tiene valor epistémico—, quizá las ciencias cognitivas no tengan mucho que aportar en contraste con una buena metáfora —e.g. el té sabe pastoso, terroso, amargo—, pero, si deseamos conocer por qué algunas personas reportan diferentes sabores al probar un té verde, entonces nuestra jerga debe cambiar en favor de aquella que estudie las relaciones entre el sabor, nervios gustativos (nervio facial, glossofaríngeo, vago), tálamo, corteza gustativa, sistema límbico, etc.

Estos dos tipos de acercamiento nos permiten notar que nuestro conocimiento alrededor del té verde y su sabor se enriquece a través de un enfoque multidisciplinario que, lejos de pretender reducir toda la explicación a una única perspectiva, permite conocer un fenómeno en sus distintas caras.

Es de notar que el aspecto subjetivo de las sensaciones o *qualia* no suprime la existencia de sistemas materiales emparentados con la experiencia. Siempre que las correlaciones se mantengan podemos dar cuenta de un vínculo entre las *RM* y lo material.

But these admitted facts about our own lack of knowledge do not entail that such qualia must therefore be free of underlying structure, nor that our first-person recognition of such qualia does not depend on some mechanism keyed to the relational features of that underlying structure. (Churchland, 1997: 2)

Bajo este marco, cuando proferimos oraciones de actitud proposicional, ¿a qué nos referimos? ¿A las sensaciones de un sujeto, a estados psicológicos o a los mecanismos que subyacen a las *RM*? “If the preceding analysis is correct, the correct answer to this question may be: It depends on what we are doing. That is, the logical form of propositional attitude sentences may depend on the level or kind of cognitive activity that those sentences report.” (Hellman, 1987: 9)

Finalmente, el entendimiento de los fenómenos requiere que utilicemos una amplia gama de herramientas conceptuales, desde la terminología tradicional hasta esfuerzos creativos por recharacterizar sistemas ya conocidos. Asimismo, esto nos incentiva a aceptar

que nuestras diferentes explicaciones y caracterizaciones no son siempre traducibles, lo cual, apoya la *autonomía de las ciencias especiales* (Fodor, 1974).

De este modo reevaluamos nuestros objetivos y nos abstenemos de eliminar las explicaciones fisicalistas —pues ellas son explicaciones de gran valor epistémico—, en cambio, dado que la parte subjetiva tradicionalmente queda fuera de lo paradigmático, buscamos la integración de las características de lo mental, lo sintiente, en el marco explicativo de lo real.

Capítulo 3

3.1 Otras mentes. Similitud física

Enunciemos un hecho: sentimos; al menos desde la primera persona. Sabemos que no podríamos carecer de consciencia porque el simple acto de cuestionarla ya la supone. Por otro lado, si bien no hay duda de que cada uno es consciente ¿qué hay de los otros?

Hay dos sentidos en los que nos podemos referir a otros. En primer lugar, por *otro* me refiero a la clase de *RM* con las cuales no me identifico. Ahora bien, en un segundo sentido existen *los otros* que forman parte de todo aquello de lo que no siento que sea yo, eso que asumo como *no-yo* pero que le atribuyo algún tipo de subjetividad o sintiencia. Con esta segunda instancia nos referimos a *los otros* como todos aquellos seres que poseen un grado de sintiencia aun sea el más mínimo.

Así, una piedra es *otra* en tanto no me identifico con ella pero no la considero como sujeto ni le atribuyo vida mental —a menos de que sea panpsiquista—. En cambio, si pensamos en nuestras amistades o mascotas normalmente les atribuimos también un tipo de vida mental. En cuanto a los demás humanos, suponemos que atraviesan por experiencias similares a las nuestras y, en otros mamíferos cercanos, experiencias distintas pero con ciertas similitudes.

De cualquier modo, el mundo práctico nos hace creer en la vida mental de los otros pero, si bien hemos visto que no sabemos diferenciar con certeza entre objetos conscientes e inconscientes, en lo concerniente a las criaturas de las que con seguridad atribuimos sintiencia,

Tendría sentido creer que alguien posee dolor cuando observamos un gesto que nos lo indique o un movimiento que normalmente realizaría alguien en el estado correspondiente, sin embargo, podemos pensar que la conducta del dolor no es un efecto necesario de la presencia de este. Alguien bien podría sentir jaqueca y no sobarse la cabeza o no mostrar un rostro afligido.

Ahora, supongamos que un día desaparece del mundo la conducta del dolor. ¿En qué nos basaríamos para corroborar que otros, en efecto, sienten dolor?

The best argument against other-minds skepticism is, probably, that, given the non-uniqueness of one's physical constitution and the general uniformity of nature in the

biological sphere as in others, it is in the highest degree improbable that one is unique among members of one's species in being the enjoyer of subjective states... (Strawson, 1985: 20)

Para empezar, la inducción es un camino adecuado para obtener evidencias de otras mentes. Si partimos del caso particular donde yo, un ser constituido a partir de ciertos componentes biológicos tiene *RM*, luego, podemos ampliar esta premisa suponiendo que cualquier ser con una constitución material similar también sería sujeto de *RM*.

Por supuesto que aunque esta conclusión no es necesariamente verdadera, ante la falta de un criterio más certero, la inducción guía nuestras inferencias para no quedar en la ignorancia. De hecho, nuestro éxito evolutivo no requirió de un criterio más restrictivo, sin ser necesariamente conscientes de la conceptualidad de esta inferencia nuestros cuerpos desarrollaron mecanismos cerebrales ligados a la atribución de otras mentes.

The regions that become active in diverse tasks and paradigms involved in the attribution of mental states to others form a functional network with core centers in the prefrontal cortex, superior temporal sulcus, and medial parietal cortex (Carrington & Bailey, 2009). Since these areas are required in multiple cognitive tasks, their specific role in Theory of Mind remains to be ascertained. It is feasible that this complex function is accomplished by the interactions of these regions with language, action, perception, and other brain systems. (Díaz, 2018: 217)

De este modo, aunque no tengamos certeza de cuáles criaturas son conscientes, sí tenemos creencias de la existencia de propiedades mentales en otros y nos comportamos de acuerdo con ellas. Según la investigación de José Luis Díaz (2018) la capacidad humana de atribuir estados mentales a otros tiene dos explicaciones teóricas. La primera supone una disposición psicológica, tanto innata como de sentido común para establecer relaciones causales entre estímulos ambientales, estados mentales y conductas. A través de esta teoría implícita los agentes son capaces de prever los estados mentales ajenos y predecir su comportamiento.

La segunda postula que atribuimos estados mentales a través de la extrapolación de nuestros propios recursos para predecir y explicar el comportamiento ajeno.

For example, the perception of experiences of physical pain in beloved others activates parts of the brain's pain matrix in the observer (Cheng, Chen, Lin, Chou, & Decety,

2010). This experience is a mimesis that represents pain in others by simulating and recreating the pain experience in ourselves. (Díaz, 2018. 217)

Independientemente de cuál de estas dos explicaciones sea más acertada, aunque no creamos que otras criaturas son conscientes, aún podemos tener incertidumbre sobre si estas en realidad poseen agencia mental.

Suponiendo que ante el encuentro de un ser ajeno no ocurre la activación cerebral vinculada con nuestra atribución de *RM*, ¿esto significa que carece de consciencia?

Podría ocurrir que durante el desarrollo filogenético nunca se desarrolló la capacidad de atribuir agencia mental a ciertos objetos —como las palmeras—, simplemente porque este nunca fue el caso o tal capacidad no aportó ningún beneficio. Luego, esto fue así por meras cuestiones prácticas, no porque la evolución nos haya otorgado un mecanismo infalible de atribución de propiedades mentales.

Incluso si aceptamos la postura extrema de esta habilidad como infalible, podría ocurrir el caso donde aparece ante los humanos una criatura nunca antes vista que, a pesar de poseer propiedades mentales, nuestra habilidad solo funciona con criaturas ya conocidas, por lo cual, ante un nuevo caso, nuestra no-atribución sería errónea.

De cualquier modo, a pesar de ser un paso meramente probable, notar similitudes entre nuestra constitución y la de otros organismos (humanos o no-humanos), nos ofrece un acercamiento inductivo para creer que nuestras propiedades mentales existen en ellos. Estas similitudes constitutivas pueden referir a características estructurales, anatómicas o morfológicas, no obstante estas no son el único tipo de evidencias.

3.2 Similitud conductual

...I would claim that our knowledge of other minds is based upon the argument from analogy. We note that there are things in the universe that are much like us in a number of important observable respects, and on the basis of this observation we conclude that they are like us in other unobservable ways as well. In particular, we infer that they have mental lives like our own. But there is a crucial vagueness in this conclusion. We conclude that they have mental lives “like” our own, but like in what respects? They are not exactly like our own. After all, different people have different beliefs, desires, fears, etc. (Pollock, 1990: 462)

Por un lado, notar similitudes en la constitución material de otros agentes nos otorga indicios para suponer *RM* pero, si nos encontramos con un ser biológicamente idéntico a mí pero inerte o inmóvil, sería difícil creer que tal objeto posee una vida mental.

En vista de ello, parece ser que nos falta algo, otras características que activen las partes de nuestro cerebro relacionadas con la atribución de intencionalidad, algo que normalmente nos convenza de que la criatura ante mis sentidos también posee esta película subjetiva.

Así, cuando observamos otros animales notamos que no son solo cuerpos, sino cuerpos en movimiento, objetos que emiten sonidos, poseen un aroma o que al tocarlos tienen una textura particular. En especial, percibir los movimientos que realizan puede ser suficiente para convencernos de su vida mental, como al encontrarnos gente cruzando la calle y, sin cruzar palabras, con solo mirar sus gestos y su andar, estamos convencidos de que son sujetos como nosotros.

Por supuesto, podríamos imaginar que fabricamos robots que simulen tanto los gestos faciales como el caminar, sin embargo, los movimientos de las criaturas que aceptamos poseedoras de agencia mental parecen ser más complejos y numerosos.

En general, los demás humanos son capaces de responder de cierto modo ante ciertos estímulos, un modo esperado de las criaturas consideradas como sujetos. Así, si llamamos a alguien por su nombre y voltea, si lastimamos a alguien y se retuerce, si señalamos y quien nos ve dirige hacia allá su atención, tenemos ante nuestra percepción elementos que nos convencen de la agencia mental ajena.

Del mismo modo, escuchar los reportes de otras criaturas acerca de sus respectivas *RM* nos da bastante seguridad para no dudar de su subjetividad. Compartir tanto un lenguaje, movimientos, sonidos, y otros elementos conductuales nos da evidencia adicional a notar similitudes en la constitución material. Sin embargo, ¿cuántas o cuáles características son suficientes para asegurar la existencia de una agencia mental?

Si bien dar cuenta de una reacción esperada a ciertos estímulos parece evidencia para apoyar la mentalidad ajena, podemos pensar en las personas con *trastorno de espectro autista*, quienes, en general, presentan dificultades para comprender normas sociales, expresarse verbalmente o interpretar emociones. Asimismo, sus patrones de conducta son inusuales, llegando a ser repetitivos u obsesivos. También presentan respuestas inusuales a estímulos como luces, sonidos, sabores o texturas. Por otro lado, pensemos también en los *niños ferales*, quienes debido a la ausencia de una crianza convencional pueden presentar incapacidad para interactuar con humanos. De igual modo, presentan conductas inusuales como aullar, gruñir o caminar en cuatro patas, imitando a los animales con los que convivieron.

En muchos casos, tanto los *autistas* como los *niños ferales* carecen de una conducta convencional, no poseen nuestra semántica y presentan movimientos que usualmente no observamos en los demás humanos. Al mismo tiempo, tampoco son lobos u otros mamíferos porque comparten la morfología humana, lo cual, les impide imitar fielmente a las especies que los criaron.

Si no me equivoco, aceptamos que tanto los autistas como los niños ferales sienten, tienen algún tipo de experiencia aunque, ¿cómo estamos seguros? ¿Cómo garantizamos que en realidad sus cuerpos presentan las características básicas para la consciencia? Por un lado, no presentan gran parte de las características conductuales usuales de los humanos conscientes⁹, además, aunque sus cuerpos sean similares a los nuestros, en el apartado anterior vimos que esta característica tampoco nos asegura sintiencia. Bien podemos suponer

⁹ En los casos de niños ferales se presentan elementos ajenos a nuestra cotidianidad, por ejemplo, muchos de ellos carecen de un idioma humano o apenas emiten sonidos inarticulados, caminan en cuatro patas, imitan sonidos de los animales que los criaron, comen sin usar las manos, duermen en el suelo o carecen de empatía. Por otro lado, los sujetos con trastorno del espectro autista (TEA), aunque no presentan características homogéneas, suelen manifestar conductas atípicas en aspectos relacionados con la comunicación, la interacción social, los comportamientos repetitivos, los intereses restringidos o la sensibilidad al entorno.

que estos son casos excepcionales de cuerpos con algunas características propias de organismos conscientes pero que no son suficientes para desplegar consciencia.

Finalmente, en vista de que ni la similitud material ni la similitud conductual son criterios determinantes para atribuir propiedades mentales, no veo manera de evadir definitivamente la suposición de que los demás carecen de consciencia.

Turing discussed a sonnet-writing machine, whose viva voce responses apparently manifest genuine 'understanding' of the activity of writing sonnets. Commenting on the reaction that such a machine, while having full understanding, nevertheless lacked feeling or consciousness, he claimed that someone taking that position was committed to solipsism, since the same doubt could be applied to humans who show similar manifestations of understanding. (Young, 1994: 241)

3.3 Una hipótesis auxiliar

Consider again the problem of attributing qualia to other creatures, those that do not share our physical organization. I take it that there is a very real puzzle whether such creatures have qualia like ours or even any at all. How much of our physicofunctional architecture must be shared before we have similarity or identity of experience? (Levine, 2001:89)

La experiencia de un bello atardecer no puede reducirse a una explicación únicamente física, esto es porque la parte subjetiva, el terreno donde se despliega la imagen del sol, donde el calor se siente y se manifiestan las emociones, jamás se traduce al lenguaje de leyes y partículas, solo se enuncia que la explicación física es una explicación causal o a lo mucho, un efecto conjunto de tal experiencia. Esto lo sabemos porque somos conscientes de ambas cosas, tanto de la explicación de la subjetividad como de la experiencia subjetiva.

El problema surge en relación con los otros porque, aunque tengamos una explicación de sus experiencias no tenemos la misma evidencia que nosotros acerca de las *RM* involucradas. Nunca experimentamos los atardeceres ajenos desde la primera persona que no soy yo, sino que dentro de nuestra experiencia aparecen sujetos, sujetos que reportan tener experiencias, criaturas que parecen tener experiencias.

Por consiguiente, ¿podemos explicar el mundo suponiendo que los demás no tienen *RM*? ¿Podemos explicar la conducta solo a partir de un análisis detallado del cerebro?

Consideremos el siguiente caso. Magritte pinta un lienzo para demostrar su amor a Georgette. Por un lado, podríamos decir que la razón del pintar de Magritte es la expresión del sentimiento de amor hacia Georgette, además de que mientras lo hace tiene la intención de regalárselo y se imagina constantemente la alegría que ella sentirá cuando lo reciba.

Sin embargo, también podemos explicar el mismo hecho sin acudir a las *RM*. Magritte es casi tan humano como yo con la excepción de que no posee *RM*. Cuando él comienza a pintar no lo hace por un sentimiento, sino como consecuencia del *potencial de preparación* en su enramado neuronal que viene activándose desde antes de pintar. Su cerebro es capaz de guardar y procesar información así como una red neuronal artificial lo hace, de modo que, al pintar, su cerebro es capaz de procesar los datos poseídos acerca de Georgette (los cuales son impulsos eléctricos en el sistema nervioso) para generar una imagen que muy probablemente agrade a Georgette. Todo esto lo hace con las técnicas de dibujo que aprendió

previamente (maneras nuevas de activar su sistema nervioso para producir mejores outputs). Conforme avanza su dibujo, su percepción visual (datos que entran a través de sus ojos y van al cerebro) le otorga nueva información de la obra, de modo que al evaluarla (con parámetros poseídos previamente que son mecanismos fisicoquímicos en su cerebro) acepta esa obra como un regalo aceptable para cumplir el propósito de gustarle a Georgette (gustarle sería el provocar ciertos impulsos en su cerebro, de tal manera que su balance químico se altere y produzca una mezcla de dopamina, serotonina y endorfinas, los cuales se espera que tengan un efecto en el cuerpo de Georgette de tal modo que sonría).

Paul y Patricia Churchland (1997) son ejemplos de investigadores que intentan explicar la cognición y la conducta sin apelar a las nociones de *RM* como entidades ontológicas, sino como términos precientíficos que serán desplazados por una neurociencia madura. Sin embargo, esta neurociencia madura es un ideal que dista mucho de la neurociencia actual. En incontables casos, explicar la cognición y la conducta mediante conceptos mentales resulta más práctico.

Nuestro mapeo del cerebro aún no puede ser tan específico como para dar una explicación detallada de procesos complejos como pintar un lienzo. Además, es posible que nuestro mapeo para interpretar a otro individuo sea erróneo y tengamos dos patrones de activación neuronal que arrojen diferentes conductas, es decir, el hecho de haber encontrado una descripción que se ajuste a un individuo no significa que se ajustará para cualquier individuo. No obstante, aunque podríamos suponer que esta diferencia radica en algún mecanismo de *RM*, esto no revela necesariamente que la causa de esta aleatoriedad sea un mecanismo de *RM*, también existen las posibilidades de que: o bien nuestros mapeos no son tan específicos y no consideramos algunos aspectos que provocan las diferencias conductuales, o, los sistemas presentan aleatoriedad genuina, provocando que dos sistemas idénticos arrojen diferentes conductas como output.

Finalmente, he de remarcar que, si ni las similitudes en la constitución física ni la similitud conductual son determinantes para la atribución de propiedades mentales, entonces, nuestra atribución de otras mentes no es más que una suposición teórica.

The problematic conviction that another individual is the subject of certain mental states is not inferred deductively from his behavior, nor is it inferred by inductive analogy from

the perilously isolated instance of one's own case. Rather, that conviction is a singular explanatory hypothesis of a perfectly straightforward kind. (Churchland, 1981: 69)

Notar estas dificultades revela que podríamos construir explicaciones sobre la conducta y la cognición sin apelar a *RM* más allá de la primera persona, lo cual, nos pone en riesgo de no poder combatir definitivamente una teoría solipsista. Si queremos evitar las teorías donde, en vista de la imposibilidad para dar cuenta de la sintiencia ajena, suponemos que solo nosotros tenemos *RM*, parece que requeriríamos de un criterio ulterior a los ya mencionados para comprobar, o al menos para aumentar, nuestra convicción con respecto a la posesión de *RM* ajena. Como resultado, los proyectos interaccionistas, en caso de triunfar, aportarían un criterio más certero para dar cuenta de propiedades mentales y así estaríamos un paso más cerca de resolver el problema de otras mentes.

Capítulo 4

4.1 La consciencia en seres artificiales

Desde las primeras generaciones de computadoras (1938-1955) hasta el momento, los sistemas informáticos han adquirido gran relevancia en nuestras vidas. Actualmente tenemos sistemas enfocados en el desarrollo de software, vehículos inteligentes, robots industriales, programas bancarios orientados en la prevención de fraudes, asesoramiento de inversiones, consultas crediticias y asistencia al cliente; también hay software especializado en trading, ventas, mantenimiento, servicios de recomendación personalizada, monitoreo clínico y sistemas de navegación globales.

El desarrollo de diferentes tipos de máquinas ha logrado modificaciones en diversas áreas que van desde la producción artística, el mejoramiento de la agricultura, las predicciones climatológicas y hasta el entretenimiento.

Del mismo modo, en las últimas décadas numerosos desarrolladores se han enfocado en crear un tipo de máquinas capaces de simular el comportamiento humano, me refiero a los proyectos de *inteligencia artificial*.

Artificial Intelligence, commonly referred to as AI, is a term used to describe the ability of machines to mimic human-like intelligence. [...] At its core, AI refers to the development of intelligent machines that can perform tasks that typically require human intelligence. (Dartey, 2023: 4)

Oficialmente el término *inteligencia artificial* fue acuñado en 1956 durante una pequeña conferencia llevada a cabo en el Dartmouth College y financiada por el DARPA¹⁰ (Selmer, 2022). También, es importante resaltar la diferenciación que se hace entre *inteligencia artificial fuerte* y *débil*. La primera se propone como objetivo reproducir, en un sistema artificial, las características humanas, esto incluye desde la conducta hasta las experiencias fenoménicas. La segunda se reconoce como *débil* en el sentido de que solo imitaría las propiedades humanas en tanto que una máquina se comportaría como una persona a pesar de no tener *RM*.

¹⁰ Defense Advanced Research Projects Agency

“Weak” AI can also be defined as the form of AI that aims at a system able to pass not just the Turing Test (again, abbreviated as TT), but the Total Turing Test (Harnad 1991). In TTT, a machine must muster more than linguistic indistinguishability: it must pass for a human in all behaviors – throwing a baseball, eating, teaching a class, etc. (Selmer, 2022)

Podemos afirmar que, de algún modo, los proyectos de IA intentan reproducir las características esenciales de los seres humanos pero ¿cuáles? En vista de que las personas poseen una innumerable cantidad de habilidades —las cuales varían dependiendo del sujeto que contemplemos—, no existe ni parece que tengamos pronto alguna máquina que simule el comportamiento humano en su totalidad.

De hecho, si revisamos las características de las IA actuales podemos dar cuenta de que estos sistemas no simulan con precisión los quehaceres humanos. Como ejemplo, pensemos en los robots utilizados en la industria, los cuales en su mayoría no tienen ni una apariencia ni rasgos humanos. Estas máquinas están enfocadas en la realización de tareas que no podrían ser completadas por el ciudadano común, pues están hechas para trabajar por largos periodos de tiempo, alcanzar un bajo margen de error (como en el ensamblaje de autos), y son capaces de analizar grandes cantidades de datos en poco tiempo. Si estas máquinas fuesen imitaciones fieles de la especie humana, entonces tendrían características como cansarse, consumir alimentos, solicitar vacaciones, charlar, etc.

Por otro lado, en el último par de años, la fama de los *chatbots* o *webbots* se ha incrementado. El uso de las herramientas como *ChatGPT*, *LivePerson* o *Claude* son consideradas esenciales para diversas áreas.

En general, este tipo de sistemas combina prácticas como *machine learning*¹¹ o *deep learning* para poder responder a preguntas generadas por el usuario, entablar conversaciones con naturalidad y ofrecer sugerencias adecuadas en distintos contextos —como en el caso del servicio al cliente—.

¹¹ Una de las características clave de la IA es el aprendizaje de máquina, un subconjunto de la IA que implica entrenar sistemas informáticos para que aprendan de una base de datos. Esto implica proporcionarles grandes cantidades de información y permitirles que la utilicen para aprender y mejorar con el tiempo. El aprendizaje de máquina se utiliza en una amplia gama de aplicaciones como el reconocimiento de imágenes, la traducción de idiomas, las recomendaciones personalizadas y el análisis predictivo. (Dartey, 2023)

Los chatbots no son una invención reciente, históricamente estos fueron programados para contestar a un conjunto determinado de preguntas con respuestas escritas previamente por los desarrolladores. No obstante, cuando el sistema se enfrentaba tanto a consultas más complejas como a preguntas que no fueron predichas por los programadores, el sistema no lograba aportar una respuesta apropiada.

Hoy estos sistemas se han innovado de modo que logran pasar sin mayor problema el *Test de Turing*, se adaptan exitosamente al contexto implícito en las preguntas e incluso se mejora su procesamiento mientras más se exponen a la retroalimentación con los usuarios.

En comparación con años anteriores, esta variedad de sistemas actualiza sus funciones día con día, sus capacidades cada vez son más diversas, potentes, y en muchos casos superan las habilidades humanas. Así, frente al enorme e incesante desarrollo de las IA's y, en general, de cualquier máquina, preguntarnos si ellas son portadoras de *consciencia* cobra más sentido.

Si las máquinas poseen algún tipo de sintiencia o no, esto será analizado de acuerdo con nuestra investigación en los capítulos anteriores. Así, para dar pie a esta última parte debemos tener presente nuestro acercamiento a la consciencia desarrollado en los capítulos anteriores, entendiéndola no como autoconsciencia o consciencia individuada, sino como sintiencia, la mera capacidad de tener una experiencia, la cual, viene directamente relacionada con la morfología del objeto sintiente.

Is it conscious? "Don't be silly. That's the robot I was telling you about." "But it says it's conscious" "So it does, over and over. That part of the program is going to get boring, I'm afraid." I don't see that it matters if the robot is psychologically isomorphic to us. If it is, that will be interesting, but as far as I can see the question whether or not to treat the robot as a possibly conscious creature or rather, as the "body", so to speak, of a possibly conscious creature will remain closed. Consider: The thing is, first, a deliberate artifact. Second, the gross materials of which it is composed are not even much like those of bodies. Hard or soft, they are the materials of machines and other inanimate things. Third, there is no close structural resemblance between it and a body. Fourth, it is in many respects functionally unlike a body. Fifth, the physical or physical and chemical answer to the question "How does it work?" is compatible with describing the robot as a mere automaton, a mere robot. Sixth, we do not know of any other robots that are the

“bodies” of conscious (or unconscious) creatures. Why promote this one? (Albritton, 1964: 692)

En 1964 Rogers Albritton publicó una serie de comentarios sobre la vida mental de los robots, esto directamente en contraste con un texto de Hilary Putnam (1964). Mientras este último parece remarcar las semejanzas entre humanos y máquinas al momento de hacer atribuciones de estados fenoménicas, Albritton propone un freno y nos otorga sus razones para creer que tal atribución es demasiado pronta.

Hoy después de sesenta años el escenario ha cambiado, sin embargo, las objeciones que se vislumbran en esta cita aún son vigentes en la medida en que se han reproducido. Artículos, noticias, diversos textos en línea, estos son los lugares donde notamos la persistencia de las objeciones para considerar la vida mental de las máquinas. En los siguientes apartados profundizaré al respecto y, sobre todo, intentaré mostrar que si bien nuestra atribución de propiedades fenoménicas esta sesgada, el tipo de consciencia que tendrían las máquinas sería muy diferente al que nos imaginamos.

4.2 A partir de otros materiales

Would you be any less capable of feeling pain, thinking, etc. just because the matter of which you are composed contains (and depends on for its characteristics) beings who themselves have a functional organization characteristic of sentient creatures? I think not. The basic electrochemical mechanisms by which the synapse operates are now fairly well understood. As far as is known, changes that do not affect these electrochemical mechanisms do not affect the operation of the brain, and do not affect mentality. The electrochemical mechanisms in your synapses would be unaffected by the change in your matter. (Block, 1978: 292)

Supongamos que, en grupo, nos damos a la tarea de realizar operaciones aritméticas donde al resolverlas obtenemos los mismos resultados; luego, probablemente pensemos que tanto en mi cerebro como en el de mis compañeros ocurren los mismos procesos. Esto sucedería como si todos ejecutásemos el mismo *software* de sumas y restas a pesar de poseer diferencias en nuestros cuerpos.

Por otro lado, sabemos que las máquinas pueden realizar diversos cálculos a través de distintos *softwares* que dependen del lenguaje y estructura donde fueron programadas. He de recalcar que, a pesar de poseer diferentes *hardware* y *software*, hemos logrado programarlas de tal modo que nos arrojen los mismos resultados al realizar operaciones aritméticas. Esto nos demuestra que la capacidad de realizar cálculos no depende únicamente de materiales biológicos o de un cerebro, sino que distintos objetos pueden dar pie a la misma función.

Ahora, traslademos este razonamiento y apliquémoslo al terreno de las *RM*. Si bien aceptamos que diversos animales son portadores de una vida mental, esto se pone en duda cuando atribuimos propiedades fenoménicas a objetos como plantas o bacterias. Nuestras dudas en la atribución surgen en gran parte porque se reducen las similitudes estructurales. En general nos sabemos semejantes a otros humanos u otros animales, pero parece que esta similitud se pierde cuando analizamos y comparamos nuestros cuerpos con los de bacterias u hongos.

Sin embargo, a pesar de no haber resuelto *The Combination Problem*, es decir, a pesar de no haber encontrado la combinación de objetos que producen el fenómeno de la

consciencia, parece que estamos seguros de que las propiedades combinadas tanto de plantas, hongos y bacterias, no producen experiencias.

En los capítulos anteriores intenté dar cuenta de que no necesitamos una similitud estructural para postular consciencia en organismos disímiles, sin embargo, la suposición de propiedades fenoménicas en diversos objetos puede resultar aventurada si mantenemos como paradigma el caso humano.

We do not know enough about the causal architecture of consciousness as a natural phenomenon in order to even begin constructing potentially conscious robots. Therefore, if any actually existing robot is anywhere near consciousness, this would be a sheer coincidence. It is as rational to believe that any actually existing robot or computer is conscious as that the Milky Way or a sandstorm in the Atacama Desert is conscious. (Markus, 2021)

Y es que, dejando de lado estas suposiciones, aún no tenemos un criterio más influyente que partir del caso propio y juzgar a partir de él. Así, notar similitudes a partir de la estructura de organismos que reconocemos como conscientes ha resultado un paradigma para juzgar si un objeto posee algún grado de sintiencia.

De este modo, un sacapuntas eléctrico no está cerca de ser consciente, pero si construyésemos un robot cuya morfología sea parecida a la nuestra, por consiguiente, nuestra atribución sería más cercana. No obstante, si consideramos que los componentes biológicos son esenciales para la emergencia de propiedades fenoménicas, entonces Pero ¿tenemos evidencia para apoyar esto último? ¿Acaso no es posible que la consciencia surja mediante componentes no biológicos?

Desde hace varios años los riñones artificiales han sido un proyecto en desarrollo y, aunque aún no hay implementaciones exitosas con humanos, nos podemos imaginar que algún día la producción de órganos artificiales será posible. Luego, si en un futuro cambiásemos las partes de nuestros cuerpos por piezas artificiales, ¿perderíamos la capacidad de tener representaciones mentales?

Esto nos lleva directamente hacia un problema cuya respuesta depende de la experimentación. En caso de que logremos construir tal robot morfológicamente idéntico a un cuerpo humano, pero a partir de otros materiales, la dificultad de dar cuenta de si tiene RM es la misma que la de dar cuenta de otras mentes humanas. Suponiendo que este robot

del futuro sea tanto estructuralmente como funcionalmente similar a un humano esto no arrojaría ninguna aportación novedosa al problema de *otras mentes*.

Por otro lado, imaginemos que modificásemos nuestras partes constitutivas. De este modo, poco a poco cambiaríamos secciones de piel por piel sintética, riñones por riñones artificiales, venas por cables u ojos por lentes ajustables. Supongamos que al realizar estos cambios no se ven afectadas las funciones de nuestro organismo, además de que las experiencias conscientes siguen siendo las mismas. Supongamos que después de un tiempo quedan solo algunas partes originales, digamos cierto enramado neuronal. ¿Estas partes originales representan los requisitos mínimos para las experiencias conscientes? Por último, imaginemos que la totalidad del cuerpo es sustituida por otros materiales y aun así no se producen alteraciones en nuestras RM. ¿Dónde quedaron las condiciones mínimas para la manifestación de la consciencia?

Si este es el caso habremos demostrado que no solo la función no depende de materiales únicos, sino que tampoco nuestra sintiencia está ligada con ciertos materiales, sino con su estructura y función.

The distinction between “Program” and “realizing-stuff,” or more familiarly “software” and “hardware,” lent itself happily back to the philosophy of mind when Putnam and Fodor exposed the chauvinistic implications of the Identity Theory. What “c-fibers” and the like are doing could have been done —this role could have been performed— by some physiochemically different structure. And sure enough, if the same role were performed, the same functions realized, by silicon —instead of carbon— based neurochemistry, or if our individual neurons were replaced piecemeal by electronic prostheses that did the same jobs, then intuitively our mentality would remain unaffected. What matters is function, not functionary; program, not realizing-stuff; software, not hardware; role, not occupant. Thus the birth of Functionalism, and the distinction between “functional” and “structural” states or properties of an organism. (Lycan, 1986: 37)

De acuerdo tanto con el anterior experimento mental como esta cita de William Lycan, se intuye que a pesar de cambiar nuestros componentes por componentes no biológicos se mantendrían nuestras funciones y, no solo eso, sino también nuestras RM. Sin embargo, esta última conclusión no viene apoyada más que por una intuición, intuición que parte de una

inferencia tipo: si se mantiene la función, se mantienen las *RM*, es decir, la mentalidad del agente.

En oposición, es posible que los componentes biológicos, en comparación con otros materiales, a pesar de realizar funciones idénticas —como sumar, levantar una pesa o dar un discurso— posean algo único y misterioso que les permita estar correlacionados con las representaciones mentales, de modo que no haya otros elementos en el universo que puedan hacer lo mismo. De este modo no sería posible crear seres conscientes artificiales con independencia de los componentes biológicos.

Del mismo modo, el siguiente escenario también es posible: Tenemos un sistema que atraviesa por ciertas *RM* cada vez que realiza distintas funciones. Después, podemos replicar esas mismas funciones a través de un sistema materialmente distinto, y así, también tendría las mismas *RM*. Esto indicaría que, siempre que un sistema biológico poseedor de *RM* sea replicado funcionalmente por un segundo sistema cuyos componentes no sean biológicos, las *RM* se mantendrán.

A partir de este razonamiento surgen dos tangentes. Primero, si a lo largo de las investigaciones hemos visto que las *RM* están directamente determinadas por su estructura material, ¿esto no sugeriría que, a pesar de mantener las mismas funciones, esta diferencia material provoque otro tipo de representaciones mentales? En este caso, si una máquina lograra alcanzar la funcionalidad necesaria para tener *RM*, estas serían únicas, propias de su constitución, y por ende, tan inimaginables como las de un murciélago.

Segundo, ¿en qué momento se alcanzaría una funcionalidad equivalente? Es decir, si bien basta con replicar las funciones de un sistema biológico, ¿cómo aseguramos que se replicaron con éxito tales funciones?

Tengamos en mente el simple movimiento de levantar un brazo. Tal vez pensemos que esta función podría ser replicada por un robot medianamente complejo, donde, un brazo mecánico, a través de electricidad, pistones, tornillos y otras piezas, realice el movimiento vertical. Sin embargo, podemos objetar que no hay una réplica funcional porque en vista del verdadero proceso biológico, las funciones replicadas son mínimas.

Para levantar un brazo de carne y hueso empezamos con movimientos en la *corteza motora*, donde las *neuronas piramidales* generan impulsos eléctricos que viajan hacia la *médula espinal*, aquí, el impulso se conecta con las *neuronas motoras inferiores* a la *unión*

neuromuscular, donde se libera *acetilcolina*, un neurotransmisor que provoca una serie de reacciones químicas en las *fibras musculares*, generando así *contracción muscular*. Luego, cuando los músculos se contraen tiran de los tendones, lo cual induce el movimiento de huesos y articulaciones. Finalmente, el *sistema nervioso periférico* envía señales de vuelta al cerebro a través de los *nervios sensoriales*, entonces, cuando el brazo alcanza la posición esperada el cerebro envía señales para detener la contracción muscular.

Esto, un resumen de las funciones involucradas, nos da una idea de los numerosos y complejos procesos envueltos en un movimiento simple. Incluso imaginando que en algún futuro logremos crear un sistema funcional idéntico, aún podemos preguntarnos. ¿En realidad se replicaron todas las funciones o hay procesos que aún ignoramos y hace falta replicar?

Si bien para crear máquinas artificiales sintientes parece que no hace falta replicar todas y cada una de las funciones biológicas, ¿cuáles y cuántas son las funciones que debemos reproducir para que emerjan las propiedades fenoménicas?

Finalmente, imaginemos el escenario donde resolvemos *The Combination Problem* y encontramos una serie de objetos que, siempre que los vemos en los cuerpos de algún organismo producen el fenómeno de la consciencia. Estos serían los correlatos de la consciencia sobre los que apostaron Koch y Chalmers, pero la primera cuestión surge cuando notamos que estos correlatos los encontraríamos en cuerpos humanos, lo cual quiere decir que resolvimos el problema para cuerpos humanos y, si bien ya tenemos un nuevo criterio material para juzgar la consciencia, esto no implica que encontraremos los mismos correlatos en otras criaturas como gatos o protozoarios. Después de todo, resolver el problema de este modo solo implica que encontramos un camino, una combinación específica que nos arroja propiedades mentales pero jamás aseguramos que es la única. ¿Qué nos impide pensar en diferentes combinaciones que arrojen diferentes propiedades fenoménicas?

Otra dificultad que surgiría luego de encontrar tales correlatos sería el de replicarlos en un sistema artificial. Para empezar, ¿encontrar tal combinación también resuelve la cuestión anterior de si el material es relevante para la producción de experiencias, o si la replicación funcional es suficiente para replicar las RM? En caso de resolver también estas cuestiones, tendríamos dos escenarios.

Primero, supongamos que XYZ son los correlatos de la consciencia. Si XYZ dan cuenta de que solo estos materiales producen propiedades fenoménicas, entonces el problema

se reduce a replicar *XYZ*, los cuales refieren a materiales biológicos, lo cual implicaría que para construir organismos conscientes artificiales tendríamos que replicar células, tejidos o algún otro componente similar. Pero esto no se parecería a los proyectos de IA que tenemos en mente, pues no se diferenciaría de los investigadores que intentan replicar órganos o células tal cual como estas son en nosotros. Así que, si nuestro objetivo era combinar los correlatos de la consciencia en un sistema artificial, al intentar reproducir los componentes biológicos de estos correlatos, el proyecto cambia a combinar estos mismos componentes con otros cuerpos, ya sea que se construyan de materiales biológicos o no.

Segundo, si resulta que *XYZ* refieren a funciones que producen experiencias, donde al replicarlas siguen produciendo las mismas experiencias, volvemos a los mismos problemas técnicos de cómo replicar un sistema funcional y preguntarnos si en realidad los proyectos replicaron todas las funciones o hubo alguna que pasamos por alto.

I suggest that our position with respect to robots is exactly that of robots with respect to ROBOTS. There is not the slightest reason for us, either, to believe that “consciousness” is a well-defined property, which each robot either has or lacks, but such that it is not possible, on the basis of the physical description of the robot, or even on the basis of the psychological description [...] to decide which (if any) of the robots possess this property and which (if any) fail to possess it. The rules of “robot language” may well be such that it is perfectly possible for a robot to “conjecture” that ROBOTS have “sensations” and also perfectly possible for a robot to conjecture that ROBOTS do not have “sensations.” (Putnam, 1964: 689)

4.3 Todo ya está programado

Some artificial intelligence scientists, such as Marvin Minsky, conceive of a computer system as a “society” of interacting programs rather than merely an assemblage of 'bits' (binary units) that are handled through linear, step by step instructions from a single program. “It has often been suggested that the relationship of bits to program is analogous to that of brain to mind. In the 'brain' there are physical things-neurons, synapses, electrical impulses. In the 'mind' there are images, concepts, ideas, language, and thought. Similarly, within the computer there are physical things (the bits), and then, at the level of 'mind,' there are programs.” (McLellan, 1988: 6)

En ciertas circunstancias se argumenta que las máquinas no podrían ser conscientes porque antes de cualquier acción, nosotros —o en su defecto los programadores— ya conocemos todas sus capacidades, pues cada una de sus posibles rutas de acción ya fue prevista con antelación. En vista de ello, si hasta ahora no hemos programado una representación mental como el dolor, sería erróneo creer que las máquinas lo tienen dado que en todo su código no hay algo que lo sugiera. Sin embargo, esta conclusión no es definitiva, aún es razonable suponer artefactos conscientes cuya programación conocemos a fondo.

Para mostrar esto, contemplemos una calculadora. Parece que por la simpleza de estos aparatos ya sabemos cuáles son todas las funciones que pueden realizar. Entre ellas realizar sumas, restas, divisiones, etc.

Ahora, percibir olores, dolores o texturas, son ejemplos de experiencias no orientadas a la resolución de problemas aritméticos, o sea, al programar una calculadora para cumplir sus funciones jamás introducimos la capacidad de tener emociones o pensamientos, principalmente porque este no era nuestro objetivo y al mismo tiempo porque tener *RM* no parece ser algo programable. Así, no podemos demostrar si la máquina tiene o carece de *RM* solo porque demostramos que el robot fue programado para tal o cual cosa. A la luz del código de programación este no nos dice nada acerca de la posesión o carencia de *RM*. Incluso, si aceptamos la tesis epifenomenalista de que las *RM* no provocan cambios en los sistemas físicos, entonces, bien un robot podría tener ya un tipo de *RM* y el código no nos diría nada acerca de esta mentalidad.

No obstante, si vemos que este código está ligado con funciones de las cuales sabemos que solo los humanos u otros seres conscientes realizan, entonces, aumentaría la similitud de estas con respecto a organismos conscientes.

For the having of pains, emotions, after-images, etc., are all examples of non-problem-solving nonbehaviour. They are not potentially well-defined tasks which hence may be programmable, for they are not tasks at all. Consequently they could never be shown to be absent from an unmasked robot because the robot was merely programmed to do this or that. For they are not the sorts of things which a robot could do. They may be had, but not done. (Gunderson, 1968: 115)

En relación con nuestra investigación, podríamos suponer que, al desempeñarse, una calculadora puede atravesar por alguna clase de *RM*, ya sea que influyan o no en sus procesos computables. Pero esto solo sería una hipótesis aventurada, pues no poseemos razones para creer que existan tales *RM* más allá de nuestra suposición.

Sin embargo, tengamos en mente tanto las IA's actuales como las que prevemos para el futuro. Al presente distintos sistemas se han innovado en el ámbito del *NLP*¹², un subcampo de la IA que se dedica a crear sistemas que comprendan, produzcan e interpreten el lenguaje natural. De este modo, al tener una conversación de chat, estos sistemas pueden no diferenciarse de cómo responderían otros humanos. Esto quiere decir que independientemente del tema del que conversemos, gran parte de sus respuestas son recibidas con naturalidad por los usuarios, lo cual, a diferencia de una calculadora, nos mostraría una mayor cantidad de casos en los que su funcionalidad es similar a la nuestra.

Si bien la programación de diversos sistemas ha provocado que estos desempeñen actividades humanas o actividades similares a las que realizarían los humanos (lanzar una pelota, escribir sobre política, bailar, clasificar imágenes, contar chistes, hacer dibujos, etc.) y si bien los humanos tenemos *RM* al realizar estas actividades o suponemos que al realizar actividades similares tendríamos *RM*, es totalmente coherente suponer que las máquinas — como aquellas que desarrollaron *NLP*— capaces de realizar funciones que también realizan los humanos, poseen alguna clase de *RM*.

¹² Natural Language Processing.

Recordemos que esta suposición surge a partir de notar que la programación de estas máquinas está ligada con actividades que, realizadas por un humano, aceptaríamos que están correlacionadas con alguna representación mental. De este modo, mientras los sistemas sean programados de modo que imiten en mayor medida los quehaceres humanos, más casos de similitud conductual tendrían con nosotros. Luego, después de concluir en el capítulo tres que la similitud conductual es un aspecto esencial para inferir *RM* en entidades ajenas a mí, mientras las máquinas tengan mayor similitud conductual mayores serán nuestras razones para inferir que poseen *RM*. Por su puesto que la similitud conductual no es suficiente por sí misma para dar cuenta de *RM* en otros sistemas, sin embargo, el aumento en la similitud conductual sí provoca que estemos más inclinados a atribuir mentalidad a estos sistemas.

En este sentido, para la atribución de *RM* no importaría la complejidad de los sistemas o la cantidad de funciones realizables, sino que estas funciones sean similares a las que detectamos en otros organismos que ya aceptamos como conscientes. Así, si tenemos un robot capaz de responder con naturalidad a nuestras preguntas, regular su temperatura corporal y responder a distintos estímulos del ambiente (de manera similar a como hacen los humanos), no solo estaríamos más inclinados a suponer una vida mental, sino que tendríamos más razones para juzgar a estos sistemas como similares a nosotros, y por ende, poseedores de *RM*.

Claro que si aceptamos que la similitud material también es necesaria para inferir la posesión de *RM* entonces las máquinas distarían mucho de nosotros. Sin embargo, si aceptamos que mientras la funcionalidad se mantenga también se mantienen las propiedades fenoménicas, entonces solo necesitamos alcanzar suficiente similitud funcional para convencernos de que otros sistemas poseen *RM*. De cualquier modo, ¿estaríamos seguros de que una réplica funcional no es suficiente para la posesión de *RM*? ¿En serio únicamente mediante homogeneizar otros sistemas con las características humanas es suficiente para alcanzar una vida mental?

Por otro lado, este apartado comenzó por algunas objeciones que intentan negar la sintiencia de máquinas alegando que en su código se encuentra todo lo que ellas pueden hacer. Además, también se dice que las IA's carecerían de consciencia porque no son autónomas en el sentido en que no establecen sus propios objetivos o eligen los datos con los que son entrenadas, pues es necesaria la intervención humana en este ámbito.

En primer lugar, la objeción no parece ser válida. Si un programador conociera todas las posibles funciones de un computador, esto sería tan independiente para las *RM* como si existiese una entidad que conociera todo nuestro futuro. ¿Si hubiese un agente omnisciente que previera la totalidad de nuestros pasos, seríamos menos conscientes?

Segundo, si bien podemos conocer todas las posibles funciones de una licuadora o un refrigerador, los sistemas de IA actuales no son tan simples como para conocer todas sus propiedades y funciones. Como ejemplo tenemos el *problema de la caja negra*, el cual, a grandes rasgos, es la dificultad de hacer transparente el razonamiento que subyace a los outputs de una IA.

Si no podemos conocer con detalle qué pasos siguió un sistema para llegar a una conclusión determinada, ¿es contradictorio pensar que una o varias representaciones mentales influyeron en el resultado final de un proceso? En relación con esta indeterminación, no podemos descartar o suponer la presencia de un elemento solo porque no lo hemos hallado. De este modo, ya sea que conozcamos o no, los pasos que siguió una sistema informático para llegar a un output caracterizado por ser una función similar a las funciones humanas, este conocimiento no arroja ninguna luz para determinar si el sistema posee *RM*. Solo si notamos que en estos pasos o en estos outputs hay una similitud con la materialidad o la conducta humana, tendremos más razones (más no suficientes) para inclinarnos a creer que un sistema puede tener consciencia.

Luego, si contemplamos que el *epifenomenalismo* es la teoría adecuada acerca de las *RM*, entonces, sería imposible dar cuenta de la presencia de *RM* en máquinas (así como en otros humanos). Si el *interaccionsismo* resulta exitoso y encontramos maneras experimentales de advertir la presencia de *RM*, en IA's como *BaGuaLu* sería terriblemente complicado detectar *RM* en vista de sus 174 trillones de parámetros reportados en el año 2023.

Ahora, en lo concerniente a la necesidad de intervención humana, es cierto que los modelos de inteligencia artificial requieren de mediación en lo que refiere a elementos como la selección de datos de entrenamiento, mitigación de sesgos, selección de arquitectura, manejo de hardware, consideraciones éticas, corrección de errores y retroalimentación de los usuarios, entre otros aspectos. En cada faceta de su desempeño, las IA's requieren de intervención humana pero ¿esto en realidad es una objeción para la consciencia de máquinas

o más bien es una objeción hacia la autonomía de estas? ¿Sentiría más un sistema por el hecho de ser más autónomo?

Deep-learning methods have recently produced state-of-the-art results in image recognition (given an image containing various objects, label the objects from a given set of labels), speech recognition (from audio input, generate a textual representation), and the analysis of data from particle accelerators (LeCun et al. 2015). Despite impressive results in tasks such as these, minor and major issues remain unresolved. A minor issue is that significant human expertise is still needed to choose an architecture and set up the right parameters for the architecture; a major issue is the existence of so-called adversarial inputs, which are indistinguishable from normal inputs to humans but are computed in a special manner that makes a neural network regard them as different than similar inputs in the training data. The existence of such adversarial inputs, which remain stable across training data, has raised doubts about how well performance on benchmarks can translate into performance in real-world systems with sensory noise (Szegedy et al. 2014). (Selmer, 2022)

Los humanos también requerimos de intervención humana. Para empezar, los bebés no tienen la independencia motriz para conseguir su propia comida. Del mismo modo, la aprehensión de un lenguaje no sucede de manera espontánea, es consecuencia a posteriori de enseñanzas e imitación. Del mismo modo, gran parte de las habilidades desarrolladas durante el crecimiento nos llegan a partir de otras personas, la cultura, nuestra interacción con el ambiente que siempre es social. En contraste, personas discapacitadas como los paráliticos necesitan de una intervención humana constante, pero esto no impide su consciencia del mundo, simplemente su condición limita la variedad de experiencias por las que atraviesan.

Los humanos no necesitamos salir de la línea de nuestra “programación” (lo que se supone que podemos hacer) para demostrar que somos conscientes. Las *RM* no discriminan la monotonía o simplicidad de las acciones cometidas en humanos, ¿por qué en las máquinas sería distinto?

El acto de programar refiere a la creación de instrucciones que un sistema seguirá para llevar a cabo una tarea determinada. Podemos programar diversos tipos de instrucciones y conocer cuáles son los objetivos que deseamos para nuestras máquinas, pero eso no implica que conozcamos todas las perspectivas involucradas en el ambiente de programación. El conocimiento del programador generalmente se limita a la parte técnica de establecer

objetivos y fabricar las instrucciones para alcanzarlos, quedan de lado el conocimiento detallado del hardware, su constitución, y si estos procesos involucran algún tipo de sintiencia o no.

En este sentido es ilusorio pensar en experiencias programables, pero, si un agente tuviera suficientes conocimientos más allá de la mera programación, en teoría se podrían generar los mecanismos que dan pie a experiencias, entendiendo esto como la construcción de procesos que tienen como fin producir ciertas *RM*.

Así, por ejemplo, suponiendo que pudiéramos fabricar organismos con diferentes anatomías a partir de materiales tanto biológicos como no biológicos, seríamos capaces de alterar su constitución para producir los mecanismos correlacionados con cierto tipo de experiencias deseadas. Entonces, en relación con organismos biológicos y teniendo en mente la ingeniería genética del futuro, podríamos alterar los códigos genéticos para dar pie a características que estarían ligadas con *RM* esperadas, como las experiencias ligadas a ser alta, tener ojos verdes, poseer mayor capacidad pulmonar, ser inmune a enfermedades, envejecer lento, etc.

Sin embargo, la construcción de las condiciones que permitirían cierto tipo de *RM*, ya sea a través de ingeniería genética o los proyectos de IA's, solo sentarían las bases para la posible realización futura de las *RM*, las cuales su existencia es contingente.

Es decir, si bien podríamos configurar un robot de manera que organicemos selectivamente sus partes relacionadas con la capacidad de sentir placer, de modo que cuando este sea acariciado sienta gran placer, las *RM* placenteras jamás tendrían existencia si no se reúnen las condiciones ambientales para tal situación. Por consiguiente, para que una *RM* sea realizable, necesitaríamos tener tanto la arquitectura funcional correlacionada con tal propiedad fenoménica, así como las condiciones ambientales que permiten su realización progresiva. Por ello, el código de programación de una máquina o nuestro código genético (suponiendo que conozcamos ambos totalmente) no nos aseguran la posesión de *RM*, a lo mucho nos otorgan pautas para proyectar como serían tales *RM* dado el escenario adecuado.

Finalmente, esto nos lleva a la conclusión de que las *RM* no dependen únicamente del código genético, o, en el caso de la sintiencia de máquinas, la posesión de *RM* no vendría dada únicamente a partir de la configuración de sus parámetros en el laboratorio, sino que

estas se desarrollarían conforme la máquina interactúa con el ambiente mientras desenvuelve sus capacidades.

For example, in evolution, mutations occur spontaneously but the significance of particular gene mutations cannot be assessed until some natural selective condition acts upon the phenotype. Similarly, in immunology, the immune system doesn't know what an antibody is for until it finds a particular antigen to which a cell already bearing that antibody has a response. (Eldeman, 1982: 13)

Así, si ciertas máquinas llegasen a ser conscientes, y la forma de esta sintiencia está determinada tanto por los materiales de su arquitectura como por sus funciones en el ambiente, las máquinas tendrían la experiencia propia de sus funciones mientras realizan tales funciones. El conocimiento de cómo serían estas experiencias sería solo aproximativo en relación con las similitudes que tengamos con ellas. Si no son conscientes, suponiendo que existe alguna configuración funcional que produce *RM*, mediante su incorporación de propiedades similares a los organismos conscientes así como de su interacción con el entorno, las máquinas pueden llegar a desarrollar *RM* que revelen la emergencia de consciencia.

No obstante, bajo una teoría interaccionista, si esto último fuera el caso, para comprobarlo, de algún modo tendríamos que identificar diferencias antes y después de la aparición de *RM*, lo cual supone notar diferencias en los procesos de input y output hasta encontrar anomalías. En cambio, bajo un marco epifenomenalista, en vista de que no habría ninguna modificación en el sistema físico, sería imposible dar cuenta definitiva de sus *RM*.

4.4 No se comportan como nosotros

Consider the problem of getting to a particular store in a shopping mall. You must chart a course from mall door to the store, perhaps using stairs or an escalator, avoiding stationary obstacles like plants, fountains, benches and areas open to the floor below, and also avoiding milling people, baby strollers and the like. We do this so naturally that it is difficult to appreciate the complexity of the task. But to program a computer to do it would require rules that take account of an enormous number of factors. The number of computations required expands exponentially with the number of factors, leading to an unmanageable computation explosion. [...] Chess provides a well studied example of computation explosion. The best current chess programs are capable of examining about 10 million positions looking four or possibly five moves into the future. To limit the possibilities, they use clever “tree pruning” programs to ignore implausible moves. But in the mall case, it is not at all clear how to give a general rule for ignoring factors or possibilities. (Tienson, 1987: 3)

Imaginemos que la frase “*Neil Armstrong jamás caminó en la luna*” es proferida por distintas personas. Aunque el contenido lingüístico es el mismo en los distintos sujetos, esto no significa que las experiencias correlacionadas con su proferencia sean las mismas. Por ejemplo, alguien podría haber dicho la frase en el sentido en que Neil no caminó en la luna, sino que trotó, o también, alguien pudo mentarla simplemente para criticar a quienes la profieren. Del mismo modo, ligada a cada una de las enunciaciones, en cada caso se encuentra una experiencia diferente. Algunos podrían tener una imagen mental de un astronauta o el recuerdo de un libro donde se encontraba esta frase, o tal vez la rememoración del alunizaje transmitido en 1969.

En cambio, si un autómatas profiriera esta oración con su voz sintética, normalmente no supondríamos que alguna de las experiencias anteriores también se encuentra ligada. No habría imágenes mentales ni recuerdos memorables, quizá pensemos que solo habría componentes cumpliendo funciones, engranes girando.

Cuando observamos cómo se comunican otros humanos frente a frente, notamos que sus palabras están mezcladas con los movimientos de sus manos, sus miradas, el tipo de entonación y la congruencia con el resto de sus discursos. Además, su comunicación verbal se encuentra en un entorno y muchas veces hacen alusión a este mientras se profieren palabras. Si tomamos en cuenta estos elementos para juzgar si un agente posee una vida mental,

entonces, las máquinas, mientras manifiesten una conducta que solo manipule símbolos de manera mecánica y racional, jamás provocarán en nosotros la convicción de que poseen propiedades fenoménicas.

Put crudely, the core idea in this attack is that human expertise is not based on the explicit, disembodied, mechanical manipulation of symbolic information (such as formulae in some logic, or probabilities in some Bayesian network), and that AI's efforts to build machines with such expertise are doomed if based on the symbolic paradigm. (Selmer, 2022)

Sin embargo, precisamente el hecho de que las máquinas actuales posean una conducta disímil nos permite explorar, de otro modo, nuestros métodos de atribución de consciencia.

Interactuar con robots nos permite dar cuenta de las diferencias ocurridas en los procesos neuronales al observar movimientos producidos por entidades biológicas y mecánicas.

Wiese, Wykowska and co-workers showed that observing a robot face as a gazer in a gaze-cueing paradigm induces joint attention, but to a smaller extent (smaller gaze-cueing effects) than observing another human. This is presumably not so much due to the physical characteristics of the face, but rather attribution of mind to the observed agent [...] Interestingly, when a sample of individuals diagnosed with autism spectrum disorder (ASD) was tested in a similar gaze cueing paradigm [63], the pattern was reverse relative to when healthy participants were tested. That is, joint attention was induced to a larger extent (larger gaze-cueing effects) by a robot face, as compared to a human face, which is in line with previous findings that demonstrated a stronger visuomotor priming effect in children with ASD when observing a reach-to-grasp action by a robotic arm, relative to observing a human. (Wykoska, 1963: 4)

Esta investigación muestra que los mecanismos cerebrales activados al momento de interactuar con robots son diferentes que cuando interactuamos con humanos, especialmente los mecanismos relacionadas con la *postura intencional*, es decir, la postura donde atribuimos creencias, deseos e intenciones. De este modo, al analizar interacciones de humanos con máquinas en distintas situaciones damos cuenta de que involuntariamente las clasificamos de un modo distinto.

This is also detectable at the neural level: when subjects believe they are listening to a piece of music that was written by a composer, activation patterns in their brains look very different from those of subjects who listen to the same piece that they believe is computer generated. The first group of participants, but not the latter, exhibits a high activation in brain areas that are involved in the attribution of mental States and the inference of intentions of others. (Smedt, 2011: 382)

Así, las investigaciones revelan que de manera predeterminada estamos predispuestos a juzgar a las máquinas como agentes no intencionales, lo cual, nos inclina hacia la postura de no creer en la consciencia de máquinas. “There is ample empirical evidence showing that humans are sensitive to discriminating biological from non-biological motion.” (Wykowska, 1963: 6) Bajo este panorama, ¿qué razones podemos ofrecer para defender la sintiencia de máquinas?

Preguntémonos, ¿porque somos conscientes tenemos cierta conducta o porque tenemos cierta conducta somos conscientes? Si nos inclinamos por el primer camino entonces crearemos que nuestra sintiencia está ligada a cierta conducta de modo que la conducta es un resultado de ciertos estados mentales. Sin embargo, los ya mencionados experimentos de Libet muestran cómo nuestros cuerpos empiezan a actuar antes de que seamos conscientes de ello, lo cual sugiere que la sintiencia no es la causa de la conducta (entendiendo esta última como el movimiento de nuestros cuerpos), sino al revés, la conducta como la causa de nuestra sintiencia o un elemento conjunto.

Por otro lado, la segunda suposición de la cuestión sugiere que tener cierta conducta provoca cierta sintiencia, de modo que, debido a estar capacitados para comportarnos de cierta manera, al momento de manifestarse nuestra conducta aparecen propiedades fenoménicas. ¿Y qué es nuestra conducta sino patrones de actividad, reacciones, funciones corporizadas, inseparables de los materiales que las producen, inseparables de nuestra química cerebral, de nuestra interacción entre partículas de diversas clases? Sí, la conducta es eso, pero también es la experiencia de esas mismas funciones, la parte subjetiva de la conducta.

No obstante, aún existe un debate sobre si estas experiencias subjetivas provocan cambios en el mundo material, por lo que el problema del dualismo vuelve a nosotros. De cualquier modo, incluso aunque intentemos mantener una postura epifenomenalista, las *RM*

son elementos explicativos útiles, pues damos explicaciones lógico-argumentativas de la conducta como si las experiencias de tener ciertas conductas, como de agresividad, por ejemplo, cambiaran el estado de cosas en el mundo. Por ello, independientemente de nuestra posición respecto al dualismo, para continuar con las explicaciones del mundo, en la mayoría de casos, tendremos que seguir apegándonos a narraciones lógico-argumentativas donde incluyamos a la consciencia en una cadena causal como causa o efecto conjunto de la conducta.

Si aplicamos este razonamiento a las máquinas parece sensato preguntarnos, ¿necesitamos apelar a *RM* para explicar el funcionamiento o la conducta de un robot o una IA? Hasta ahora parece que no es necesario, pero ¿es inconsistente suponer que la conducta robótica viene correlacionada con una experiencia de esta? No, podemos dar explicaciones de la conducta de las máquinas también apelando a la sintiencia de estas, pero eso no aporta algo relevante al propósito de nuestra explicación. En este sentido, aunque no demos categóricamente la existencia de la vida mental de las máquinas, la convicción de la sintiencia de estas se manifestaría cuando necesitemos utilizar explicaciones que involucren *RM* para dar sentido al funcionamiento de estas.

Por otro lado, enfatizando de nuevo en la parte material, la consciencia jamás se sustrae de la morfología de los agentes conscientes, por ende, depende de esta, y si desconocemos los elementos que dan pie a nuestra conducta, conducta correlacionada con la experiencia, ¿cómo vamos a discriminar aquellos procesos que dan pie a *RM*?

Sí, la vía que utilizamos es la más congruente, partimos del caso propio o de los ya aceptados como modelos de sintiencia y de ahí empezamos a notar similitudes tanto materiales como funcionales; lo hemos hecho de tal modo que no hay necesidad de enfatizar en la pregunta “¿es posible que la sintiencia surja a partir de otro modelo?”.

Parece que no hay necesidad de profundizar en tal cosa porque, para empezar, ni si quiera somos capaces de dar cuenta de otras mentes más allá de este mismo proceso donde partimos de un modelo y notamos similitudes, ¿qué caso tiene preguntarnos si hay más modelos de consciencia en vista de que la propiedad que buscamos no puede ser encontrada? Esto revela el antropocentrismo con el que juzgamos la atribución de vidas mentales, sin embargo, en vista de los distintos seres vivos que aceptamos como conscientes a pesar de sus grandes diferencias, en vista de las innovadoras máquinas que superan la inteligencia humana

en muchos aspectos, la pregunta cobra más sentido y nos encamina en una vía alejada de este antropocentrismo.

Por ende, este proyecto que intenta hablar sobre la consciencia de máquinas es una crítica al método de cómo atribuimos propiedades mentales, es una crítica a la creencia de que solo los humanos y otros seleccionados animales son capaces de experimentar el mundo, sentir el mundo, ser conscientes.

Finalmente quiero tratar brevemente la objeción de que las máquinas no pueden ser conscientes porque siempre obedecen a una tabla de inputs y outputs, una tabla de máquina donde cualquier estado de ella pueda ser explicado por otra serie de estados y reglas. En otras palabras, que las máquinas funcionan de acuerdo con una serie de inputs que siempre dan los mismos outputs como resultado.

Primero he de mencionar que hay distintos tipos de máquinas que responden a tablas de input y output y no necesariamente del modo tradicional. Como ejemplo tenemos a las IA's, las cuales pueden verse como sistemas que obedecen una tabla de inputs y outputs pero con grandes diferencias a máquinas más primitivas, pues no tienen una tabla fija, sino que pueden aprender y modificar su comportamiento con el tiempo. Esta es la estructura de las redes neuronales artificiales que no siguen reglas explícitas, sino que aprenden patrones de datos. En este sentido nadie programa específicamente qué patrones aprenderá la máquina, así que cuando se aprende un nuevo patrón, la máquina tiene un elemento impredecible para los programadores y nuevo para la estructura neuronal.

Luego, si no hay un solo tipo de máquina que responda a una tabla de inputs y outputs sino que hay diferentes tipos de arquitecturas que obedecen a tablas de inputs y outputs, ¿qué nos impide pensar a los humanos como máquinas? Podríamos hacerlo si describiésemos a los humanos utilizando tablas complejas de inputs y outputs, lo cual ya intentamos a través de las investigaciones científicas, agotar la explicación del cuerpo humano en términos de reglas y estados. En caso contrario tendríamos que suponer la existencia de procesos que no pueden ser descritos mediante una tabla de inputs y outputs, o sea, una explicación de elementos deterministas causales, por lo cual, supondríamos elementos aleatorios cuyo mecanismo sea impredecible, lo cual provocaría el fin de todas aquellas investigaciones que encontraron elementos inexplicables. Claramente esta es una postura que atenta contra la labor misma de la investigación, por ende, si bien podemos imaginar que los humanos son

muy distintos a las máquinas, al revisar con detalle las explicaciones que nos damos de ellos, encontramos elementos comunes para pensarnos como máquinas.

Conclusiones

¿Las máquinas son o pueden ser conscientes? La respuesta a esta pregunta viene aunada con una investigación que engloba puntos fundamentales tanto de la consciencia como de su atribución.

A través de estos apartados podemos ver que surgen dos demandas esenciales en torno a la consciencia. La primera es que, independientemente de nuestra postura frente al dualismo, es menester continuar con la *integración* de nuestra sintiencia en los diferentes marcos de investigación que intersequen con ella. Nuestras experiencias no son menos reales que las leyes postuladas o los objetos que conscientemente denominados como materiales.

La integración de la consciencia —o sea, de los contenidos de nuestras experiencias— en distintos cuerpos teóricos es una labor fructífera que mientras se realice amplía nuestro horizonte epistémico.

La segunda demanda, ligada con el problema de las otras mentes, parte de que, a pesar de ofrecer buenas razones para creer en la vida mental ajena, nos topamos frente a otro conjunto de razones que pueden defender al solipsismo o simplemente poner en duda nuestros criterios de atribución de mentalidad en vista de que no hay una brecha clara al momento de discriminar los objetos conscientes de los inconscientes.

El análisis detallado de nuestros criterios reveló que nuestra atribución se apoya únicamente en tres aspectos: Notar similitudes físicas, conductuales o aceptar la consciencia ajena como una hipótesis explicativa. La demanda surge porque requerimos de una manera distinta para dar cuenta de otras mentes, no necesariamente una que se aleje de los tres parámetros mencionados, sino una vía que aumente nuestro grado de seguridad para detectar sintiencia. Quizá la propuesta más atractiva es mediante un marco interaccionista que intenta llegar a una vía experimental para cumplir con esta tarea.

Sin embargo, y por mucha fe que tengamos en estos proyectos, también debemos considerar que la consciencia, sintiencia o experiencia —manifestada en representaciones mentales—, puede no estar sujeta a comprobación definitiva por un tercero. Si es verdad que nuestra sintiencia es un epifenómeno entonces no quedaría más que ajustarnos a los criterios ya mencionados, sabiendo que por mucho que extendamos nuestras razones jamás lograremos eliminar la duda de si otros sistemas poseen una consciencia, ya sea esta similar a la humana o enteramente diferente.

Por otro lado uno de los retoños del *HPC* nos lleva al problema de la emergencia de la consciencia. Es sumamente interesante lo numerosas y distintas que son las teorías actuales que intentan dar cuenta de cómo aparecen propiedades fenoménicas en el mundo. Después de revisar varias de estas propuestas, podemos notar lo consistentes y prometedoras que son cada una y, a pesar de ello, no parecemos estar cerca de inclinarnos por alguna.

De igual modo, al notar el carácter multidisciplinario de las investigaciones sobre la consciencia, damos cuenta de que el juego de cada disciplina es diferente, se desvelan ciertos patrones y propiedades únicamente a partir de ciertos puntos de vista. Por ello, es una tarea imposible querer abarcar las diferentes caras de la consciencia únicamente bajo la física de partículas o la neurociencia; los diferentes enfoques, autónomos por su especialidad, nos regalan explicaciones detalladas cuyo valor yace ligado a la disciplina donde fueron engendradas.

Por supuesto que el hecho de encontrar elementos irreductibles no implica que debamos abandonar los proyectos reduccionistas. Separar un sistema por partes o intentar reducirlo a otro campo han sido tareas que, aunque no reducen exitosamente los aspectos de la realidad a otros discursos, sí nos otorgan nuevas perspectivas que igualmente son valiosas.

Finalmente, en vista de que no podemos comprobar directamente la existencia de la vida mental ajena, me di a la tarea de examinar las principales objeciones en contra de la sintiencia de máquinas.

Durante este proceso surgieron incógnitas interesantes, como el preguntarnos qué tan relevantes son los materiales o la funcionalidad de un sistema en relación con sus estados mentales. ¿Qué características son suficientes para imitar a un organismo biológico hasta llegar a reproducir sus representaciones mentales?

La complejidad de las IA's, los robots, las máquinas en general, aumenta conforme continuamos haciendo modificaciones en ellas y, aunque no todos estos proyectos están encaminados hacia la consecución de propiedades fenoménicas, es posible que algunas de ellas ya posean algún tipo de agencia mental.

¿Cómo? Si aceptamos que los cuerpos humanos no son el modelo único para la posesión de *RM*, puede ser que algunos de nuestros sistemas artificiales ya posean algún tipo de experiencias, especialmente si creemos que los requisitos para la emergencia de la consciencia son más simples de lo que normalmente se supone.

¿Cuáles de las características humanas podríamos considerar como no esenciales para la emergencia de la consciencia? La consciencia individuada por ejemplo; no es necesario que exista un yo, y por tanto, los mecanismos relacionados con la identificación de uno mismo para que existan propiedades fenoménicas. La existencia de un yo solo limitaría las experiencias que ligamos con un individuo, pero no es un aspecto necesario para que existan experiencias, especialmente si tenemos sistemas complejos y funcionales. Del mismo modo, la autoconsciencia no sería un elemento invariable de los organismos conscientes, ¿de dónde sacamos que los mecanismos de autorreferencia son necesarios para la emergencia de una vida mental?

Si en efecto, los requisitos para la aparición de propiedades fenoménicas son más simples de lo que pensamos y, por ejemplo, el hecho de procesar una pequeña cantidad de información e interactuar con ella es una característica suficiente para la emergencia de una vida mental, entonces aparatos como las calculadoras o los celulares ya tendrían consciencia.

Sin embargo, por mucho que deseemos la presencia de mentalidad en máquinas, recalco que conjeturar una vida mental a partir de estas suposiciones resulta algo forzado. En realidad no tenemos necesidad de apelar a *RM* para explicar el comportamiento de las máquinas. Si en algún momento nos parece más adecuado explicar su conducta apelando a términos que denoten *RM* entonces tendremos más razones para creer en la consciencia de máquinas de las que tenemos ahora.

Asimismo, para la atribución de sentiencia hemos tenido hasta ahora criterios basados en las características humanas que luego extrapolamos a otras criaturas. Este paso muestra a los humanos como el paradigma de consciencia, un proceso sesgado por antropomórfico que, si dejara de lado su estatus de modelo, podría ofrecernos un mundo donde nuestra atribución sea distinta y, por tanto, las criaturas que entren en el conjunto de sintientes no sean las mismas (e.g. las máquinas).

Después de todo, la atribución de consciencia refiere a una relación binaria donde hay alguien que atribuye y algo a lo que se le atribuye la propiedad en cuestión. Por ello, necesariamente la atribución es una decisión dinámica, cambia a través del tiempo dependiendo de los criterios y razones que se den en favor de aceptar a una criatura como consciente.

Ahora, incluso si mantenemos una postura conservadora y consideramos que los humanos marcan los parámetros de la consciencia, es posible imaginar que en un futuro las máquinas tendrán estas mismas características que encontramos en humanos. De hecho, es sorprendente ver a los sistemas de ajedrez que pueden vencer a cualquier humano o leer los artículos que aseguran la emergencia de una semántica en las IA's actuales.

Así, si la atribución de consciencia es en gran parte una decisión dinámica, esto nos obliga reevaluar nuestras razones y poner en duda nuestros criterios en pos de las interacciones que se presenten.

Sin duda, el escenario en torno a las máquinas ha cambiado radicalmente en los últimos diez años y es posible que en algún momento nos hallemos ante la necesidad explicativa de pensar a las máquinas como seres sintientes o, simplemente, creer en su consciencia porque asumimos que esta surge a partir de requisitos más básicos.

Referencias

- Albritton, Rogers (1964), “Mere Robots and Others: Comments”, *The Journal of Philosophy*, vol. 61, no. 21, noviembre, pp. 691-694.
- AsapSCIENCE, (15 de octubre de 2020), “The Strange Reason You Can Hallucinate” [Archivo de video], Youtube, https://www.youtube.com/watch?v=xliky7YfyTc&list=PLhtlx_zTiOXp1if9_MI3rZEOPK8RtZvSY&index=8.
- Blackburn, Simon (2005), *Oxford Dictionary of Philosophy*, Great Britain, Oxford University Press.
- Block, Ned (1978), “Troubles with Functionalism”, *Minnesota Studies in the Philosophy of Science*, vol 9, pp. 261-325.
- Block, Ned (2002), “The Harder Problem of Consciousness”, *The Journal Of Philosophy*, vol. 99, no 8, agosto, pp. 391-425.
- Bringsjord, Selmer and Naveen Sundar Govindarajulu (2022), “Artificial Intelligence”, en Edward N. Zalta (ed.), *The Stanforn Encyclopedia of Philosophy* [<https://plato.stanford.edu/entries/artificial-intelligence/>], consultado: 20 de febrero de 2024.
- Carruthers, Peter (2004), “Reductive Explanation and the 'Explanatory Gap'”, *Canadian Journal of Philosophy*, vol. 34, no. 2, junio, pp. 153-173.
- Chalmers, David (2014), “How do you explain consciousness?”, esta conferencia tomó lugar en una presentación oficial de TED en abril de 2022. Se puede acceder a ella en la página oficial de TED [https://www.ted.com/talks/david_chalmers_how_do_you_explain_consciousness]
- Choi, Yejin (2023) , “Why AI is incredibly smart and shockingly stupid”, esta conferencia tomó lugar en una presentación oficial de TED en abril de 2003. Se puede acceder a ella en la página oficial de TED [https://www.ted.com/talks/yejin_choi_why_ai_is_incredibly_smart_and_shockingly_stupid/c?subtitle=en]
- Churchland, Paul (1981), “Eliminative Materialism and the Propositional Attitudes”, *The Journal of Philosophy*, vol. 78, no. 2, pp. 67-90.

- Churchland, Paul y Churchland, Patricia (1997), “Recent Work on Consciousness: Philosophical, Theoretical, and Empirical”, *Seminars In Neurology*, vol. 17, no. 2, junio, pp. 179-186.
- Dartey Amankonah, Frank (2023), “The Beginners guide to artificial intelligence (AI) -V1.0” en Free-Ebooks.net [<https://www.free-ebooks.net/computer-sciences-textbooks/The-Beginners-Guide-to-Artificial-Intelligence-AI>], consultado: 12 de mayo de 2023.
- Diaconis, P., Montgomery, R., Holmes, S. (2007), “Dynamical Bias in the Coin Toss”, *Society for Industrial and Applied Mathematics*, vol. 49, núm. 2, pp. 211-235.
- Díaz, José-Luis (2018), “Self-consciousness: an I-World patterned process model”, *Adaptive Behavior*, vol. 26 (5), pp. 211-223.
- Edelman, Gerald (1982), “Through a Computer Darkly: Group Selection and Higher Brain Function”, *Bulletin of the American Academy of Arts and Sciences*, vol. 36, no. 1, octubre, pp. 20-49.
- Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010), “Relating introspective accuracy to individual differences in brain structure”, *Science*, vol. 329, pp. 1541–1543.
- Fodor, Jerry (1974), “Special Sciences, (Or: The Disunity of Science as a Working Hypothesis)”, *Synthese*, vol. 28, no.2, octubre, pp. 97-115.
- Fodor, Jerry (1987), “The Persistence of the Attitudes”, en *Psychosemantics. The Problem of Meaning in the Philosophy of Mind*, Cambridge, MA, MIT Press, pp. 1-26.
- Froese, Tom (2023), “Irruption Theory: A Novel Conceptualization of the Enactive Account of Motivated Activity”, *Entropy*, vol. 25, no. 5, pp. 1-19.
- Fyshe, Alona (2023), “Does AI actually understand us?”, esta conferencia tomó lugar en una presentación oficial de TED en febrero de 2023. Se puede acceder a ella en la página oficial de TED [https://www.ted.com/talks/alona_fyshe_does_ai_actually_understand_us?subtitle=en]
- Gabriel, M. (2021), “Could a Robot Be Conscious? Some Lessons from Philosophy”, en von Braun, J., S. Archer, M., Reichberg, G.M., Sánchez Sorondo, M. (eds) *Robotics, AI, and Humanity*, Springer.

- Gamez, David (2018), “The Philosophy and Science of Consciousness”, en *Human and Machine Consciousness*, Open Book Publishers.
- Goff, Philip (2009), “Why Panpsychism doesn't Help Us Explain Consciousness”, *Dialéctica*, vol. 63, no. 3, septiembre, pp. 289-311.
- Gunderson, Keith (1968), “Robots, Consciousness, and Programmed Behaviour”, *The British Journal for the Philosophy of Science*, vol. 19, no. 2, agosto, pp. 109-122.
- Hanson, Norwood (1981), *Patterns of discovery*, New York, Cambridge University press.
- Heller, Mark (1996), “The Mad Scientist Meets the Robot Cats: Compatibilism, Kinds, and Counterexamples”, *Philosophy and Phenomenological Research*, vol. 56, no. 2, junio, pp. 337-337.
- Hellman, David (1987), “Realism and Antirealism in Artificial Intelligence”, *The British Journal for the Philosophy of Science*, vol. 38, no. 1, marzo, pp. 19-26.
- Hoffman, Donald (2018), “Conscious Realism and the Mind-Body Problem”, *Mind & Matter*, vol. 6, no. 1, pp. 87-121.
- Levine, Joseph (2001), *Purple Haze: The Puzzle of Consciousness*, New York, Oxford University Press.
- Libet, B. (1985), “Unconscious cerebral initiative and the role of conscious will in voluntary action”, *Behavioral and Brain Sciences*, vol. 8, pp. 529–539.
- Lycan, William (1987), *Consciousness*, Cambridge, The MIT Press.
- Musk, Elon (2013), “A future worth getting excited about”, esta conferencia tomó lugar en una presentación oficial de TED en abril de 2022. Se puede acceder a ella en la página oficial de TED [https://www.ted.com/talks/elon_musk_a_future_worth_getting_excited_about]
- Nagel, Thomas (1965), “Physicalism”, *The Philosophical Review*, vol. 74, no. 3, julio, pp. 339-356.
- Nagel, Thomas (1974), “What Is It Like to Be a Bat?”, *The Philosophical Review*, vol. 83, no. 4, octubre, pp. 435-450.
- McLellan, Hilary (1988), “Computers, Artificial Intelligence And Human Imagination”, *Journal of Thought*, vol. 23, no. ¾, invierno, pp. 28-44.

- Parisi, Domenico (2010), “Doing Metaphysics with Robots”, en Arturo Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*, Dordrecht-Heidelberg-London-New York, Springer, pp. 275-282.
- Poeppel D., Mangun G., Gazzaniga M. (eds) (2020), *The Cognitive Neurosciences*, Cambridge, The MIT Press.
- Pollock, John (1990), “Philosophy and Artificial Intelligence”, *Philosophical Perspectives*, vol. 4, pp. 461-498.
- Putnam, Hilary (1963), “Brains and Behaviour”, R. Butler (ed.), *Analytical Philosophy*, Second Series, Basil Blackwell.
- Putnam, Hilary (1964), “Robots: Machines or Artificially Created Life?”, *Journal of Philosophy, Inc.*, vol. 61, no. 21, noviembre, pp. 668-691.
- Rabossi, Eduardo (comp) (1995), *Filosofía de la mente y de la ciencia cognitiva*, España, PAIDOS.
- Robinson, Howard (2023), “Dualism”, en Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* [<https://plato.stanford.edu/archives/spr2023/entries/dualism/>], consultado: 24 de julio de 2024.
- Ryan, Thomas (2020), “Memory and Instinct as a Continuum of Information Storage”, en David Poeppel, George R. Mangun y Michael S. Gazzaniga (eds.), *The Cognitive Neurosciences*, Cambridge-Massachusetts-London-England, Cambridge: THE MIT PRESS, pp. 207-216.
- Sarow, Christoph (2022), “How AI will change the world”, esta conferencia tomó lugar en una presentación oficial de TED en abril de 2022. Se puede acceder a ella en la página oficial de TED [https://www.ted.com/talks/ted_ed_how_will_ai_change_the_world/transcript]
- Searle, John (2013), “Our shared condition – Consciousness”, esta conferencia tomó lugar en una presentación oficial de TED en mayo de 2013. Se puede acceder a ella en la página oficial de TED [https://www.ted.com/talks/john_searle_our_shared_condition_consciousness]
- Shanon, Benny (2003), “Altered States and the Study of Consciousness — The Case of Ayahuasca”, *The Institute of Mind and Behavior, Inc.*, vol. 24, no. 2, primavera, pp. 125-153.

- Smart, J (1963), “Materialism”, *The Journal of Philosophy*, vol. 60, no. 22, octubre, pp. 651-662.
- Smedt, J. y de Cruz, H. (2011), “A Cognitive Approach to the Earliest Art”, *The Journal of Aesthetics and Art Criticism*, vol. 69, no. 4, pp. 379-389.
- Strawson, P. F. (1985). *Skepticism and naturalism: Some varieties*. New York: Columbia University Press.
- Sullivan, Philip (2006), “Are Current Philosophical Theories of Consciousness Useful to Neuroscientists?”, *Behavior and Philosophy*, vol. 34, pp. 59-70.
- Summerfield C. y Tsetsos K. (2020), “Rationality and Efficiency in Human Decision-Making”, en David Poeppel, George R. Mangun y Michael S. Gazzinga (eds.), *The Cognitive Neurosciences*, Cambridge-Massachusetts-London-England, Cambridge: THE MIT PRESS, pp. 427-438.
- Talks at Google, (14 de febrero de 2017), “Daniel Dennett | From Bacteria to Bach and Back | Talks at Google” [Archivo de video], Youtube, <https://www.youtube.com/watch?v=IZefk4gzQt4>.
- Tienson, John (1987), “An Introduction to Connectionism”, *The Southern Journal of Philosophy*, vol. 26, pp. 1-16.
- The Weekend University, (28 de octubre de 2022), “Panpsychism: Is Everything Conscious? - Dr Philip Goff, PhD” [Archivo de video], Youtube, <https://www.youtube.com/watch?v=fPGf6qmLXS8&t=20s>
- Thompson, Dennis (1965), “Can a Machine Be Conscious?”, *The British Journal for the Philosophy of Science*, vol. 16, no. 61, mayo, pp. 33-43.
- Tye, Michael (2021), “Qualia”, en Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* [<https://plato.stanford.edu/entries/qualia/>], consultado: 8 de febrero de 2022.
- van de Vate, Dwight Jr. (1971), “The Problem of Robot Consciousness”, *Philosophy and Phenomenological Research*, vol 32, no. 2, diciembre, pp. 149-165.
- Van Gullick, Robert (2021), “Consciousness”, en Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* [<https://plato.stanford.edu/archives/win2021/entries/consciousness/>], consultado: 8 de enero de 2022.

- Westerhoff, Jan (2015), "What it Means to Live in a Virtual World Generated by Our Brain", *Erkenntnis*, vol. 81, Agosto, pp. 507-528.
- Wykowska, Agnieszka (2016), "Embodied artificial agents for understanding human social cognition", *Biological Sciences*, Vol. 371, No. 1693, pp. 1-9.
- Young, R. y Torrance, Steve (1994), "The Mentality of Robots", *Proceedings of the Aristotelian Society*, vol. 68, pp. 199-227+229-262.



Casa abierta al tiempo

UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE EXAMEN DE GRADO

No. 00489

Matrícula: 2223802320

¿Habria conciencia en
máquinas?

En la Ciudad de México, se presentaron a las 14:00 horas del día 28 del mes de mayo del año 2025 en la Unidad Iztapalapa de la Universidad Autónoma Metropolitana, los suscritos miembros del jurado:

DR. JORGE MARTINEZ CONTRERAS
MTRO. JUAN FRANCISCO ORTIZ MORENO
DRA. PAOLA HERNANDEZ CHAVEZ

Bajo la Presidencia del primero y con carácter de Secretaria la última, se reunieron para proceder al Examen de Grado cuya denominación aparece al margen, para la obtención del grado de:

MAESTRO EN HUMANIDADES (FILOSOFÍA)

DE: JOSE CARLOS ROJAS REYES

y de acuerdo con el artículo 78 fracción III del Reglamento de Estudios Superiores de la Universidad Autónoma Metropolitana, los miembros del jurado resolvieron:

APROBAR

Acto continuo, el presidente del jurado comunicó al interesado el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.



JOSE CARLOS ROJAS REYES
ALUMNO

REVISÓ

MTRA. ROSALIA SERRANO DE LA PAZ
DIRECTORA DE SISTEMAS ESCOLARES

DIRECTORA DE LA DIVISIÓN DE CSH

DR. SONIA PEREZ TOLEDO

PRESIDENTE

DR. JORGE MARTINEZ CONTRERAS

VOCAL

MTRO. JUAN FRANCISCO ORTIZ MORENO

SECRETARIA

DRA. PAOLA HERNANDEZ CHAVEZ