Aproximaciones al argumento de la habitación china de John Searle

María Alejandra Tabares Cardona

Facultad de Artes y Humanidades
Departamento de Filosofía
Universidad de Caldas
Manizales, Colombia
Noviembre de 2020

Aproximaciones al argumento de la habitación china de John Searle

Tesis de Maestría en Filosofía

María Alejandra Tabares Cardona

Director: José Fernando Ospina Carmona

Facultad de Artes y Humanidades
Departamento de Filosofía
Universidad de Caldas
Manizales, Colombia
Noviembre de 2020

Agradecimientos

Muchas personas han contribuido en mi formación académica de diferentes maneras. Considero que el apoyo emocional y moral contribuyen en buena parte en la formación de cualquier ser humano, por ello, agradezco a mi familia, especialmente, a mi madre, quién me ha apoyado en todo momento.

Mis más sinceros agradecimientos al profesor José Fernando Ospina Carmona, por dirigir y orientar este trabajo. Agradezco todas sus enseñanzas a lo largo de mi formación académica y profesional; el empeño, la dedicación y el esfuerzo por lograr transmitir sus conocimientos y de esta forma mantener la filosofía viva. También quisiera agradecer a cada uno de los docentes del Departamento de Filosofía por impartir sus conocimientos y hacer posible mi formación filosófica.

Finalmente, agradezco a aquellos amigos que se han formado conmigo y que han compartido muchas experiencias académicas que me han enriquecido. Gracias a Estefanía Moreno, Sebastián Sánchez, Martha Cecilia Meza, por todas sus enseñanzas y discusiones sobre diversos temas, incluido, el de la presente investigación.

Tabla de contenido

Introducción

En la literatura filosófica contemporánea se encuentran una cantidad de problemas que van de la mano con los avances sociales, culturales, científicos, tecnológicos, etc., se sabe que el mundo se ha revolucionado tecnológicamente y que la tecnología cada vez impacta más la vida de los seres humanos. Es por ello, que el tema de investigación del presente trabajo centra su atención en si existe la posibilidad de llegar a interactuar con máquinas en un futuro como seres semejantes a los humanos. Para ello se hará toda una reconstrucción histórica de lo que se ha definido como inteligencia artificial (IA), al mismo tiempo, se analizarán conceptos que comprometen el estudio de dicho tema, tales como mente, conciencia, pensamiento, intencionalidad, estados mentales, entre otros.

El primer capítulo de este trabajo de investigación recoge un cúmulo de opiniones sobre los primeros estudios que se hicieron en materia de IA. La pregunta que se pone en discusión y sobre la que versa toda clase de opiniones tiene con ver con la posibilidad de que las máquinas piensen. A partir de dicho planteamiento, se dan a conocer una serie de experimentos mentales que van a cuestionar la pregunta central del capítulo.

Los experimentos mentales que se mencionan, incluida la máquina de Turing y los juegos, van a ser primordiales para ofrecer una primera definición acerca de conceptos como conciencia, intencionalidad y mente, los cuáles serán rebatidos en el capítulo siguiente. En resumidas palabras, se dejan abiertos muchos interrogantes que serán tema de investigación en capítulos posteriores.

La filosofía de la mente de John Searle ha sido crucial para problematizar e intentar clarificar algunos conceptos que involucran el problema de la IA, por lo que el segundo capítulo, se enfoca en la filosofía de la mente de Searle. El problema mente-cuerpo es uno de los primeros abordajes que se hace, para poder explicar la conciencia como algo que hace parte de la naturaleza biológica, y no como algo inmaterial y diferente a la biología humana. En este sentido, se ofrece una solución por parte de Searle al tan estudiado problema mente-cuerpo, teniendo en cuenta los procesos de micronivel y macronivel; posteriormente, se abordan conceptos como los de intencionalidad, subjetividad y causación mental, todos ellos, rasgos de lo mental, según la filosofía de Searle.

En el tercer capítulo se ofrece una explicación detallada sobre el argumento de *La habitación china* de Searle, donde se retoman preguntas y planteamientos que se dejaron expuestos en el primer capítulo. Acá, la pregunta inicial "¿Pueden las máquinas pensar?" es seriamente cuestionada y rebatida, ya que para Searle la pregunta correcta debe ser "¿Puede un computador digital, tal como se ha definido pensar?" Alrededor de esta pregunta y del experimento que planteó el filósofo, se genera toda una discusión en la que se concluye que un programa de computador está definido de manera puramente sintáctica, y el pensamiento no consiste únicamente en manipular símbolos, sino también en atribuir significados, por lo que resulta pretensioso afirmar que un programa puede pensar.

También, se mencionan algunas de las objeciones y réplicas que algunos estudiosos del tema le han hecho a Searle, así como las respuestas del filósofo a las mismas. En el capítulo cuarto, estas réplicas serán ampliadas y analizadas con más detalle, con el fin de encontrar los puntos débiles en la argumentación de dicho autor.

Aunque posteriormente se ahondará en tales réplicas. Una primera réplica, es la de los sistemas, en la que se argumenta que el hombre de la habitación del experimento mental de Searle, es solo una parte del sistema, analógicamente, como la CPU, o como la unidad de procesamiento central, que tan solo hacen parte de un sistema más grande, ya que este está formado por otros componentes que incluyen una base de datos, una memoria, una serie de instrucciones, etc. De tal manera que es el sistema en su totalidad el que responde a las preguntas en chino, y no una parte como lo pretende hacer ver Searle. Una segunda réplica, es la de la mente virtual, la cual sostiene que el operador de la habitación china no comprende chino únicamente por correr o instanciar la máquina de papel. En dicha réplica se argumenta que lo importante es que la comprensión es creada. Una tercera réplica es la de las otras mentes, la cual defiende que para atribuir mente no es necesario la existencia de un cerebro. En ese sentido, se le puede atribuir mente a todo aquello que responda a los estímulos del medio y que pueda interactuar con los humanos. Una cuarta réplica es la de la intuición, en la que se sostiene que el argumento de Searle se basa en la intuición. Finalmente, se analizan dos artículos que muestran las debilidades del argumento de Searle.

Para la conclusión de este trabajo de investigación, la autora se vale de los aportes del filósofo Daniel Dennett. En un primer momento, se expone el argumento que Searle visualizó una década después de haber publicado el argumento de *La habitación china,* en el que afirma que la

sintaxis no es intrínseca a la física, sino relativa al observador; la autora se vale de este argumento para justificar que, en el mismo sentido, las creencias y deseos son relativos al observador. Posteriormente, acude al concepto de estrategia intencional propuesto por Dennett, la cual es la mejor forma de explicación de las conductas humanas.

En resumidas palabras, de hecho, no hay una cuestión que permita hacer discriminaciones sobre lo qué es un creyente, porque termina siendo algo relativo y permeado por con un contexto que podría cambiar. Incluso, teniendo en cuenta el desarrollo tecnológico, podría llegar un momento en la historia, en el que la humanidad interactuara tanto con las máquinas que olvidara que solo son máquinas.

CAPÍTULO I

Aproximaciones a la inteligencia artificial

El presente capítulo tiene como propósito fundamental ahondar en una discusión de gran interés filosófico: el surgimiento de la IA y su impacto en la vida del hombre. La pregunta "¿Puede pensar una máquina?"[1], es el tema de mayor interés a lo largo del capítulo, pero antes de enfrentarse a un problema de tal magnitud es necesario hacer una reconstrucción histórica sobre la aparición de los computadores, el propósito para el cual fueron diseñados y las distintas modificaciones que permitieron nuevos hallazgos y avances en materia de IA. En el transcurso del capítulo se revisarán algunas especulaciones acerca de si una máquina puede pensar.

Según fuentes históricas, el computador se creó en Alemania, pero se desarrolló de manera simultánea y diferente en tres países: Alemania, Gran Bretaña y Estados Unidos. El primer computador empezó a funcionar en 1941 en Alemania, su finalidad era realizar operaciones matemáticas que no requerían mayor complejidad; además, la función que cumplía era tan específica que era útil para un grupo selecto o para una actividad específica. Sin embargo, en Gran Bretaña sucede algo diferente y novedoso, si bien el propósito de estas máquinas era descifrar los códigos secretos de la Wehrmacht (códigos secretos de la época de la guerra), en 1943 aparece un computador electrónico que va a recibir el nombre de Colossus, un invento bastante revolucionario para la época, ya que su función consistía en descifrar mensajes codificados, su única limitación era realizar grandes cálculos matemáticos. El hecho de poder descifrar mensajes animaba a la creación de nuevas máquinas con técnicas cada vez más eficientes y útiles que aportaran a las necesidades de la época.

En Estados Unidos el desarrollo del computador fue más lento, aunque, terminó dominando la producción industrial informática, ya que su interés en materia de IA fue creciendo y progresando. La primera máquina que se aprobó en 1945 fue el ENIAC (Computador e Integrador Numérico Electrónico), construido en la Universidad de Pennsylvania por John Mauchly y J. Presper Eckert; su programación era bastante compleja, y para cumplir una nueva tarea debía ser reprogramado (Copeland, 1996). Por lo que en 1949 crean el BINAC, de acuerdo a Copeland

[1] Pregunta extraída del artículo "*Computing Machinery and Intelligence*" de Alan Turing.

(1996) “una empresa de Manchester, Ferranti Limited, contrató la producción de una versión comercial del Mark I de Manchester. Estas máquinas fueron los primeros computadores electrónicos de programa almacenado que se manufacturaron comercialmente en el mundo” (p. 22). Además de la aparición de computadores electrónicos, estos permitían almacenar información, algo nuevo y beneficioso para añadir a este grandioso invento. Posteriormente, crearon el UNIVAC, otro computador electrónico que revolucionó el mundo de la industria y que sería fundamental para continuar con la creación de nuevas máquinas e implementaciones que permitieran mostrar hasta donde llegaría la tecnología.

Hasta ese momento la tarea del computador no va más allá de realizar operaciones matemáticas, descifrar mensajes codificados y almacenar información, operaciones básicas para las múltiples necesidades de los seres humanos. Ahora bien, ¿en qué momento se empezó a hablar de IA? ¿Por qué hablar de inteligencia? ¿Bastará con que una máquina sea capaz de realizar algunas operaciones matemáticas y descifrar mensajes para ser considerada como inteligente? John Von Neumann fue uno de los primeros en utilizar el término de IA, quien para el año 1956 ejecutó un ordenador, un primer intento en programación de IA. Dicho programa lleva el nombre “Logic Theorist” pero ¿Por qué John Von Neumann utilizó ese nombre para su programa? ¿Qué relación tiene la lógica con la IA?

Según Copeland (1996):

> La lógica es una preocupación central de la investigación en IA. La capacidad de razonar lógicamente es, por supuesto, un componente fundamental de la inteligencia humana; si los computadores han de alcanzar la categoría de inteligencias artificiales, hay que darles la capacidad de buscar lógicamente la solución de un problema. (p. 26)

Así los programas no solo se limitan a registrar datos y realizar cálculos, sino también a crear lo realmente novedoso, un programa capaz de ejecutar su propio plan de acción. Por su parte Newell, Shaw y Simon consideraban que, si la lógica buscaba solucionar problemas, debería existir un programa capaz de inventar pruebas de teoremas de geometría elemental. En un primer momento esto no resultó exitoso, pero posteriormente se lograron hacer demostraciones de teoremas de lógica pura. Ahora el computador es una máquina capaz de “razonar” y de realizar pruebas de enunciados abstractos, lo que demostraba un avance en IA.

Aunque John Von Neumann fue uno de los primeros en interesarse por la IA, el padre de la misma fue Alan Turing, quien en su artículo titulado "*Computing Machinery and Intelligence*" formuló la siguiente pregunta: "¿Puede pensar una máquina?" Esta pregunta es central en cualquier discusión sobre IA, además involucra una discusión filosófica seria, ya que para dar una posible respuesta a este interrogante es necesario revisar qué se entiende por pensamiento y quiénes tienen la facultad de hacerlo. Cuestión que será revisada más adelante, en el análisis de conceptos.

Pero ¿Qué pasaría si de pronto usted pudiera conversar con una máquina? ¿Si esta emitiera respuestas con sentido a sus preguntas? Parry es uno de los tantos programas que se crearon con el fin simular una conversación "real", es capaz de responder a las preguntas que le hace un psiquiatra. Ejemplo de lo anterior, son estas dos preguntas realizadas por un psiquiatra y las respectivas repuestas dadas por Parry: "¿Quién te trajo aquí? LA POLICÍA. ¿Qué problemas tienes con la policía? LOS POLIS NO HACEN SU TRABAJO" (Copeland, 1996, p. 35).

Se podría estar de acuerdo con Copeland cuando afirma que "Parry es tan convincente que, cuando a alguien se le pide decidir por medio de una entrevista si se están comunicando con un programa o una persona, tiene auténtica dificultad en averiguarlo. A menudo no puede más que conjeturar" (Copeland, 1996, p. 37), el tipo de respuestas que ofrece Parry a los interrogantes del psiquiatra son coherentes. Pero ¿Responder coherentemente a una pregunta implica que haya pensamiento? Esa es la cuestión que interesa analizar acá a través con los siguientes casos, es decir, si realmente sostener una conversación implica pensar.

Eliza, la psicoterapeuta, una creación de Joseph Weizenbaum, es uno de los programas más famosos de IA, a diferencia de Parry, este programa realiza preguntas en forma de entrevista. Se puso a prueba con periodistas, psiquiatras y profesores, y todos coincidían en que no existía diferencia alguna entre un psicólogo que realiza un diagnóstico y la máquina. Sus preguntas eran tan coherentes que los entrevistados terminaban envueltos en la conversación.

> La máquina podía inducir a las personas sin esfuerzo a revelar sus más íntimos secretos, y a veces a Weizenbaum le resultaba difícil convencerlas de que Eliza no era una persona auténtica. Incluso su propia secretaria —que a cierto nivel sabía muy bien que la creación de Weizenbaum era simplemente una máquina— cogió la costumbre de obligar al resto del personal a abandonar la habitación para que ella y Eliza pudieran hablar en privado. (Copeland, 1996, p. 38)

Fue tanto el impacto que generó, que en un informe de la revista *Journal o f Nervous and Mental Disease* se consintió la posibilidad de hacer uso de la misma como una herramienta terapéutica de la que se podría disponer en centros hospitalarios y psiquiátricos para contribuir de manera eficaz en problemas concernientes a la salud mental. Esta posibilidad replanteó las especulaciones modernas acerca de reemplazar por una máquina la función que cumplen algunos seres humanos en un campo específico, en este caso, el campo de la psicología. Por su parte, Weizenbaum, el creador de Eliza, afirma que el punto álgido de la discusión no consiste en saber lo que puede o no hacer la máquina, y el fin para el cual ha sido realizada, sino que

> Él cree que las inteligencias artificiales, por su propia naturaleza, serían incapaces de comprender cabalmente la condición humana o de simpatizar con ella. Sin embargo, como demuestra la reacción ante Eliza, puede que haya una disposición excesiva a confiar a estas inteligencias extrañas la administración de los asuntos humanos. (Copeland, 1996, p. 39)

Y esto no está lejos de ser real, ya que el experimento que realizó el creador de Eliza con diferentes personalidades solo confirma que los seres humanos de alguna manera pueden simpatizar con una máquina, y que los entrevistados se sienten a gusto contando sus problemas y viendo las respuestas que arrojaba el programa. En ese sentido, es posible hablar de una relación entre inteligencias artificiales y humanas.

No obstante, afirmar que Eliza piense es algo apresurado y ambicioso, de hecho, lo que realmente hace es devolver a sus pacientes sus propios enunciados, esto es, está debidamente programada como se verá más adelante. Copeland sostiene que este programa no sabe nada de su entorno, está aislado de la cultura y, por lo tanto, de los rasgos que definen la personalidad de los sujetos que fueron objeto de experimento, no tiene la capacidad de planificar y visualizar acciones, y mucho menos tiene la capacidad de aprender y comprender. Facultades que en su mayoría son propiamente humanas y que por tal razón esta máquina no puede ejecutar de manera autónoma. Mas, el problema no se soluciona con afirmar que programas como estos no superan las facultades humanas. La investigación y experimentación en materia de IA es amplia, y después de programas como Parry y Eliza se diseñaron otros que vale la pena revisar acá.

Shrdlu[2] es un programa que controla el brazo de un robot que opera sobre una mesa llena de bloques de colores de varias formas y tamaños. A través de un teclado se establece una comunicación entre la persona que da las órdenes y las funciones que debe cumplir el robot. Este programa:

> Inventa y ejecuta su propio plan de acción. Su capacidad para interpretar instrucciones en inglés ordinario es impresionante, y se muestra hábil al conjeturar el significado más probable de una orden ambigua. El poder de razonamiento de Shrdlu le permite contestar preguntas complicadas sobre el mundo de su mesa, y puede, dentro de ciertos límites, examinar sus propios "motivos". (Copeland, 1996, p. 40)

Al parecer este programa, en su momento, era mucho más eficiente que Eliza y se acercaba más a lo que buscaba la IA: una "autonomía" por parte del programa. ¿Será que Shrdlu puede pensar? ¿Planificar acciones como el movimiento de los cubos y la clasificación de colores y formas? ¿Tendrá la capacidad de decidir? Cualquiera que observe el experimento podrá decir que sí, que este programa cuenta con la capacidad de ejecutar ese tipo de acciones como cualquier ser humano lo haría guiado por una orden. Pero esto no es cierto, el éxito del robot depende de la programación y, por tal motivo, no puede cumplir cualquier acción, solo aquellas para las cuales fue programado.

En los laboratorios de IA se percataron de esa insuficiencia, aunque la idea inicial era construir programas que pudieran ejecutar sus propias acciones. La opinión popularizada fue que, si los programas funcionaban de manera "perfecta", solo se debía a una buena programación. La mayoría coincidía en que ninguno de los programas mencionados anteriormente podía pensar y ejecutar acciones de manera voluntaria, todos están debidamente programados para cumplir con los propósitos para los cuales fueron diseñados. Sin embargo, la creación de Hacker[3] pone en duda esa opinión popular, ya que se caracterizaba por ser un "programa programador", es decir, él mismo programa programaba y ejecutaba los movimientos del brazo del robot del experimento de Terry Winograd[4]. "Es fundamental en la actuación de Hacker su Biblioteca de Técnicas de Programación, y un almacén de conocimientos sobre la escritura de programas. Gran parte de la

[2] "El nombre es una palabra sin sentido que Winograd tomó prestada de la revista MAD" (Copeland, 1996, p. 40).
[3] "Habitante del laboratorio de IA del MIT, y su especialidad es escribir programas para el computador en que opera" (Copeland, 1996, p. 44).
[4] Shrdlu, el robot.

biblioteca consta de recetas detalladas para construir programas por secciones" (Copeland, 1996, p. 44).

Con la creación de Hacker se estaba alcanzando el objetivo propuesto por la comunidad de científicos informáticos. Un programa que pudiera ejecutar ordenes gracias a un programador y no a un ser humano. Este programa ofreció otra visión sobre lo que hasta el momento se venía entendiendo como IA. De aquí en adelante se crearon una serie de juegos con otro tipo de pretensiones, aquellas que se acercaran más a la solución de problemas y a la ejecución "voluntaria".

Uno de los primeros juegos que se diseñaron fue *El programa de damas de Samuel*, el autor pretendía programar un computador para que aprendiera de la experiencia. "En su cauta forma de expresarse, el objetivo era programar 'un computador digital para que se comportara de tal manera que, si esa conducta perteneciera a seres humanos o animales, se diría que entraña aprendizaje' (Copeland, 1996, p. 47). Podía participar y obtener jugadas sin ningún margen de error, en pocas palabras, el computador era mil veces mejor que el ser humano que se enfrentaba a él. Otro juego novedoso fue el *GPS* (*El Resolutor General de Problemas*), con el que se pretendía construir un computador que buscara inteligentemente sus propias soluciones a los problemas, este juego no estaba muy alejando del objetivo principal en materia de IA. Recordando, que John Von Neumann buscaba programas que fueran capaces de resolver sus propios problemas, de ahí la necesidad de introducir la lógica a sus programas.

El GPS logró resolver problemas[5] tal y como los resolverían seres humanos que razonan, tanto así, que Newell y Simon describen este programa como un simulador del ser humano, esto es, como un programa que pueda razonar de manera similar a los seres humanos. Si el GPS simula conductas humanas y es un programa de computador, se sigue que el GPS "es un indicio llamativo en favor de la hipótesis de que 'la conducta libre de un ser humano pasablemente inteligente se puede entender como el producto de un conjunto complejo, pero finito y determinado'" (Copeland, 1996, p. 53). Esta apreciación determina al ser humano como ser predecible, en la medida en que también puede cumplir ciertas funciones, al igual que una máquina.

[5] Pequeños acertijos y juegos que requerían del raciocinio.

Frente a este tipo de programas que se acaba de analizar existe una tesis generalizada que sostiene que un computador no podrá ser tan inteligente como la persona que lo diseñó, además, solo puede cumplir órdenes. Esto parece incuestionablemente verdadero, toda vez que se ha visto que el éxito de programas como Eliza, Parry y Shrdlu depende únicamente de la programación que un ser humano hace. Mas, una de las conclusiones a las que llega Copeland (1996) es que esta tesis no es del todo tan cierta y que

> Un programa que aprenda de manera tan consistente a vencer a la persona que lo diseñó está, por definición, haciendo movimientos mejores que cualquiera de los que esa persona le puede enseñar a hacer. Lo que Samuel consiguió fue proporcionar a un computador instrucciones que, al final, le permitieron hacer más de lo que él era capaz de decirle que hiciera. (p. 48)

Estos juegos han demostrado que pueden cumplir más órdenes de las que se le programan, que de alguna manera actúan de manera inteligente, entendiendo inteligente como la capacidad de ejecutar por sí mismo una acción.

Es necesario volver a la pregunta que compromete a la filosofía con la IA: "¿Puede pensar una máquina?" Cualquiera que haga una apreciación a la ligera sobre los programas se acaban de mencionar, diría que sí. Pero está claro que programas como Parry, Eliza y Shrdlu, no poseen la capacidad de comprender, aprender, interpretar información y, mucho menos, de pensar. Lo máximo que pueden hacer estos programas es cumplir órdenes. En los juegos tampoco existe algo intencional, o sea, jugadas premeditadas por la máquina. Según Copeland (1996):

> Las atribuciones de una mente a las máquinas actuales son pura metáfora. Decir que el computador ajedrecista piensa que le estoy atacando su torre de rey es una manera cómoda de expresarse, y ciertamente se deja pronunciar mejor que otras alternativas más literales. (p. 64)

Quienes definitivamente se niegan a creer que una máquina piensa o en un futuro pueda llegar a hacerlo, suelen relacionar el pensamiento con la consciencia, así, todo aquello que piense debe ser consciente, una máquina no es consciente, por lo tanto, una máquina no piensa. Pero ¿Será válida esa premisa de que todo aquello que piense debe ser consciente? Copeland quiere mostrar que esa premisa es falsa, y que además es una concepción errada, ya que el ser humano no es consciente de todos sus procesos mentales. Para ilustrar se mencionarán algunos ejemplos: la

visión ciega[6] demuestra que la percepción no tiene que ser una actividad consciente. Sucede que a un paciente se le extirpa una parte de la corteza visual del cerebro y pierde su campo visual, se le pide que cuando escuche un sonido toque un punto de la pantalla, cuando se termina el experimento, el paciente se da cuenta que inconscientemente acertó todas las veces que debía tocar el punto. Pero este no es el único caso de actividades inconscientes. En muchas ocasiones se asocian las palabras de forma consciente, pero no siempre se hace así. Puesto que se tiene la facultad de pensar, pero aun así no se es consciente de todos los estados mentales, lo que definitivamente niega esa premisa.

Turing describe un experimento[7] de laboratorio con el que posiblemente se encontrará una respuesta a la cuestión filosófica de la IA (pensamiento y consciencia). En su artículo "Maquinaria computacional e Inteligencia", menciona algunos argumentos a favor y en contra de la posibilidad de hallar pensamiento y consciencia en una máquina. Tales argumentos permitirán esclarecer la discusión acerca de la pregunta inicial mencionada en este capítulo "¿Puede una máquina pensar?"

Este científico propone *El juego de la imitación*, sobre el cual versa toda una discusión acerca de las funciones que puede cumplir una máquina y, desde luego, un ser humano. Por lo tanto, se empezará por describir en qué consiste. Para este juego son necesarios tres participantes, un hombre (A), una mujer (B) los cuales cumplen el papel de ser interrogados, y un interrogador (C). Todos tendrán una comunicación por teclado, para que así el interrogador no llegue ligeramente a la conclusión de quién es el hombre y quién es la mujer. La idea central del juego consiste en que el interrogador descubra, a través de preguntas, quién es el hombre y quién es la mujer. Los sujetos que están siendo interrogados deben engañar al interrogador para no ser descubiertos fácilmente. Si C le pregunta a A algo acerca de su apariencia física, este debe intentar engañarlo para que no descubra que es un hombre.

Lo que se pretende es que una máquina pueda responder a todas las preguntas que hace el interrogador, convenciéndolo de tal manera, que no dude de que quién responde a sus preguntas es un ser humano y no una máquina. En párrafos anteriores, se mencionó que Parry era un programa tan convincente que muchos se interesaron por sostener conversaciones con él, incluso, preferían dialogar sobre sus problemas emocionales con la máquina y no con un psicólogo. No

[6] La visión ciega fue descubierta y nombrada por Larry Weiskrantz y sus colaboradores a principios de los años setenta.
[7] Conocido como prueba o test de Turing.

obstante, no resulta sencillo afirmar que la máquina puede ser tan inteligente o más inteligente que un ser humano para persuadir a un interlocutor.

Si se piensa en las similitudes que pueden tener una máquina y un ser humano, se llega a la conclusión de que estas pueden imitar su conducta, Turing describe que un computador digital debe tener las siguientes partes: almacenamiento, unidad ejecutiva y control. Estas máquinas al igual que el ser humano almacenan información (toda aquella que se instala en la memoria humana), procesan información y realizan cálculos, por lo que se podría hablar de una imitación en la conducta humana. Pese a que, se encuentran diferencias radicales que llevan a replantear todo el problema acerca de si de las máquinas pueden pensar. Por ejemplo, un ser humano recuerda las tareas que debe cumplir, sabe que a determinada hora se debe acostar, o que debe hacer X actividad que no hizo durante el día y que hace parte de su rutina. Una máquina no puede recodar esas actividades de modo que,

> si uno quiere hacer que una máquina imite el comportamiento de un computador humano en alguna tarea compleja, se le debe preguntar cómo lo hace, y luego traducir la respuesta a una tabla de instrucciones. Construir tablas de instrucciones es usualmente conocido como "programación". "Programar una máquina para que lleve a cabo la operación A" significa poner la tabla de instrucciones apropiada en la máquina de manera tal que la realice. (Turing, 1950, p. 5)

El anterior planteamiento conduce a la conclusión expuesta en líneas anteriores, para que esas máquinas tengan éxito en el juego, o en alguna otra tarea, deben ser programadas, y dicha programación debe ser tan eficiente que sus respuestas simulen actitudes y comportamientos humanos. En este sentido, para que el juego de la imitación funcione, la máquina debe estar programada de tal forma que pueda imitar al interrogado, con la precisión y agilidad que un ser humano tiene a la hora de responder preguntas y cometer errores.

Una pregunta del tipo "¿Hay computadores digitales imaginables que tendrían un buen desempeño en el juego de la imitación?" (Turing, 1950, p. 8), podría ser respondida de manera afirmativa, pero solo si está programada de manera correcta. De hecho, Turing considera una serie de objeciones que imposibilitan la imitación. En primer lugar, la rapidez a la hora de realizar un cálculo matemático delataría al interrogado, la máquina es mucho más ágil para realizar cálculos matemáticos que la mente humana, y aunque haya casos excepcionales, no todo el mundo tiene tal

agilidad. Además, si se le programa para dar respuestas como las del juego de la imitación, algunas las respondería de manera errónea, sobre todo aquellas que necesitan una respuesta afirmativa o negativa. En pocas palabras, este tipo de réplicas alejan un poco más la posibilidad de que las máquinas piensen.

El científico tampoco es claro a la hora de definir el concepto de pensar; plantea una serie de argumentos y objeciones que muestran similitudes con el comportamiento humano, y que ponen en duda la posibilidad de que realmente haya "pensamiento". Entre los distintos argumentos que plantea Turing en su artículo "Maquinaria computacional e Inteligencia", se encuentra el del profesor Jefferson, quien sostiene que:

> Hasta que una máquina pueda escribir un soneto o componer un concierto debido a las emociones y pensamientos que tuvo, y que no sea debido al uso de símbolos al azar, podremos estar de acuerdo que máquina es igual al cerebro, es decir, no sólo que lo escriba, sino saber que lo escribió. (Turing, 1950, p. 11)

Por lo que no se trata solo de manipular símbolos o adaptar una serie de reglas, se trata de comprender semánticamente lo que se está expresando. Los reparos del profesor Jefferson van más allá de la posibilidad de establecer una manual de reglas e incluso de la propia programación. Al parecer, una máquina solo puede pensar, si esta, al menos está en la capacidad de comprender lo que se está trasmitiendo.

La objeción de Lady Lovelace es similar al argumento que presentó el profesor Jefferson, donde afirma que una máquina no hace nuevo, dicho de otro modo, repite las mismas funciones para la cual fue programada. Esta objeción responde al argumento de que como las máquinas no pueden ejecutar programas por sí solas, difícilmente pueden ser consideradas como inteligentes. Así mismo, el *Argumento de la discontinuidad en el sistema nervioso* recuerda que "si cada hombre tuviera un grupo determinado de reglas de conducta por las cuales él regulara su vida, no sería más que una máquina. Pero no existen tales reglas, así que los hombres no pueden ser máquinas" (Turing, 1950, p. 16), en otras palabras, no puede existir una comparación entre las máquinas y los seres humanos.

Una réplica a *La objeción de Lady Lovelace* versa sobre la idea de que algunas operaciones que realiza la mente son meramente mecánicas, y las máquinas cumplen funciones mecánicas, en este sentido, se podría pensar en una similitud entre la mente y la máquina. Ahora bien, tal similitud

podría llevar a hacer otro tipo de planteamientos tales como que una máquina pueda aprender, y aunque físicamente va a ser muy incompetente, si se compara con un niño humano, intelectualmente podría estar en un mismo nivel, al menos podrían ejecutar tareas muy similares.

Ahora bien, las preguntas que surgen son: ¿Pueden las máquinas realmente aprender? ¿Cuál es la condición para que se genere un aprendizaje y desde luego un conocimiento? Es evidente que las máquinas pueden desempeñar eficientemente determinados trabajos; aun cuando, la discusión acá no versa sobre las condiciones físicas de las máquinas y los resultados que pueden aportar en materia de trabajo industrial, el problema es netamente intelectual e incluso sensitivo, lo que para algunos intelectuales implica una responsabilidad moral y social.

Roger Penrose, es uno de los intelectuales interesados en el problema en cuestión, de hecho, retoma el experimento de Turing para replantear algunos argumentos acerca de la posibilidad de que las máquinas piensen o sientan. Supongan que la conversación que están teniendo el interrogador y la máquina fluye, y que el interrogador sigue convencido de que con quien está sosteniendo la conversación es con un ser humano, ¿qué pasaría cuando la máquina no comprenda lo que el interrogador quiere expresar? ¿Cuándo no haya una comprensión semántica? Seguramente, el interrogador se va a dar cuenta de que con quien ha sostenido una conversación, no ha sido con un ser humano sino con una máquina. “La habilidad de la interrogadora radicaría, en parte, en imaginar estas preguntas originales, y en parte, en hacerlas seguir de otras, de naturaleza exploratoria, diseñadas para descubrir si ha habido o no una ‘comprensión’ real” (Penrose, 1996, p. 16), o en su defecto, se podrían plantear preguntas sin sentido para detectar que tampoco es un ser humano.

¿Qué pasaría si las máquinas superan la prueba de Turing? Esta pregunta inquietó a Penrose, porque siendo ese el caso se podría hablar de pensamiento y sentimiento. Este académico está convencido de que una máquina no puede imitar al ser humano, simplemente basta con que responda eficientemente a las preguntas que se le plantean, incluso que pueda engañar a su interlocutor de manera correcta. Si todo esto se cumpliera, Penrose afirma que efectivamente se generaría pensamiento, sin necesidad de que haya toda una estructura neurobiológica que produzca el pensamiento.

Otro de los intereses de Penrose está supeditado a las funciones que pueden desempeñar las máquinas, así “los objetivos de la IA son imitar por medio de máquinas, normalmente

electrónicas, tantas actividades mentales como sea posible, y quizá, llegar a mejorar las que llevan a cabo los seres humanos" (Penrose, 1996, p. 18). Como bien se sabe, las máquinas han reemplazado muchas tareas que antes el ser humano realizaba, con una mayor eficiencia y calidad, un ejemplo de ello son las grandes máquinas industriales. Sin embargo, cuando el ser humano se enfrenta a un problema que involucra la inteligencia y el pensamiento, se niega a la posibilidad de que esto pueda ser posible, y surgen preguntas tales como las que se planteó Penrose (1996): "¿Es posible que la experiencia y competencia de los profesionales pueda ser realmente reemplazada por estos paquetes?"[8] (p. 19).

Es un hecho que diferentes programas realizan jugadas inteligentes, y son mejores que las que un ser humano puede realizar. "Las computadoras que juegan ajedrez proporcionan los mejores ejemplos de máquinas que poseen lo que podría ser considerado como 'conducta inteligente'" (Penrose, 1996, p. 20). En líneas anteriores se mencionaron las características de diferentes juegos, en especial el del ajedrez, el cual era difícilmente superado por una jugada humana, y en caso de ser superado era considerado un caso excepcional. Así mismo, se mencionaron los programas como Hacker, el cual podía autoprogramarse, lo que pone en cuestión la posibilidad de hablar de conductas inteligentes y de hacer un acercamiento a la existencia de pensamiento.

Si en este momento se concluyera con la premisa de que las máquinas no pueden pensar, la respuesta no sería estable y duradera, porque en el futuro, se contará con máquinas más veloces, con acceso más rápido, con capacidad de realizar múltiples tareas al tiempo, etc. lo que muy posiblemente lleve a afirmar que las máquinas pueden superar las tareas intelectuales del ser humano.

1.1. Inteligencia artificial fuerte

La IA fuerte defiende la idea de que las máquinas, además, de imitar los compartimientos, actitudes y respuestas humanas, están en la capacidad de pensar. Así, la actividad mental no es más que una secuencia de operaciones a la que llaman algoritmos[9].

[8] Sistema de expertos, con los que se intenta codificar el conocimiento esencial de toda profesión: medicina, abogacía, etc., un paquete de ordenador.

[9] "La palabra "algoritmo" procede del nombre del matemático persa del siglo IX Abu Ja'far Mohammed ibn Mûsa al-Khowârizm, autor de un interesante texto matemático, escrito alrededor del año 825 d.C., titulado "Kitab al jabr wa'1-

> Según la IA fuerte, la diferencia entre el funcionamiento esencial del cerebro humano (incluyendo todas sus manifestaciones conscientes) y el de un termostato radica sólo en que el primero posee una mucho mayor *complicación* (o quizá 'mayor orden de estructura' o 'propiedades auto-referentes', u otro atributo que pudiéramos asignar a un algoritmo). (Penrose, 1996, p. 24)

Desde el punto de vista de la IA fuerte, el sentimiento, la inteligencia, la comprensión y la conciencia son el resultado de un conjunto de algoritmos complejos que ejecuta el cerebro. No obstante, la complejidad del mismo impide igualarlo y, en caso de ser así, no habría ningún reparo en afirmar que las máquinas poseen pensamiento. En pocas palabras, para los defensores de la IA fuerte, no existe ninguna diferencia en que el algoritmo sea ejecutado por un cerebro o una computadora, lo realmente importante es el algoritmo.

Para comprender mejor cómo funciona el algoritmo, es importante tomar como referencia la máquina de Turing. Esta máquina no hace parte de un conglomerado de partículas que forma un artefacto, por el contrario, hace parte de un elemento matemático abstracto. En líneas anteriores se mencionó que el inventor de dicha máquina fue Alan Turing, quien para el año 1935 quiso tratar un problema matemático conocido como *Entscheidungsproblem*, planteado por el matemático David Hilbert, el cual buscaba un procedimiento algorítmico general para resolver cuestiones matemáticas.

> El problema de Hilbert que interesaba a Turing (el *Entscheidungsproblem*) iba más allá de cualquier formulación concreta de las matemáticas en términos de sistemas axiomáticos. La pregunta era: ¿existe algún procedimiento mecánico general que pueda, *en principio*, resolver todos los problemas de las matemáticas, que pertenezcan a alguna clase bien definida?. (Penrose, 1996, p. 38)

Ante dichos cuestionamientos, Turing se dio a la tarea de pensar en un dispositivo que pudiera realizar cálculos definibles en términos finitos. La idea era que dicho dispositivo tuviera un conjunto discreto de posibles estados diferentes, en un número finito, a los cuales llamó estados internos del dispositivo. Según el algoritmo de Euclides:

---

muqabala". El que en la actualidad se escriba "algoritmo", en lugar de la forma antigua, y más aproximada, "algorismo", se debe a una asociación con la palabra "aritmética". También es digno de mención que la palabra "álgebra" procede del árabe "al jabr" que aparece en el título de su libro" (Penrose, 1996, p. 34). También se refiere a "cierto tipo de procedimiento de cálculo" (Penrose, 1996, p. 24).

> No hay límite para la magnitud de los números sobre los que el algoritmo puede actuar. El algoritmo —o el *procedimiento* general de cálculo— es el mismo independientemente de la magnitud de los números. Para números muy grandes el procedimiento puede durar mucho tiempo y necesitar una gran cantidad de papel donde realizar las operaciones, pero el *algoritmo* será el mismo conjunto *finito* de instrucciones (Penrose, 1996, p. 39)

De esta manera, Turing representaba los datos externos y el espacio de almacenamiento como una cinta sobre la cual se hacen marcas. Así, el funcionamiento del dispositivo consistía en mover la cinta hacia delante o hacia atrás. Además, podía realizar nuevas marcas y borrar las que ya tenía la cinta anteriormente. De esta manera,

> La cinta consiste de una secuencia lineal de cuadros que se considera infinita en ambas direcciones. Cada cuadro de la cinta está en blanco o contiene una sola y única marca[10]. El uso de cuadros marcados o sin marcar ilustra el hecho de que estamos admitiendo que nuestro entorno (es decir, la cinta) puede ser descompuesto y descrito en términos de elementos *discretos* (y no continuos). Esto es razonable si queremos que nuestro dispositivo funcione de un modo fiable y perfectamente definido. Estamos admitiendo que el entorno sea (potencialmente) infinito como consecuencia de la idealización matemática que estamos utilizando, pero en cualquier caso particular el *input*, el cálculo y el *output* deben ser siempre *finitos*. (Penrose, 1996, p. 40)

Como se menciona en el párrafo anterior, tanto el *input*, como el cálculo y el *output* son finitos, dicho de otro modo, que tiene una capacidad limitada de información. El comportamiento de un dispositivo está determinado por el estado interno y el *input*, de esta manera:

> Hemos simplificado este *input* haciendo que sólo sea uno de los dos símbolos "0" o “1”. Dado el *input* y su estado inicial, el dispositivo actúa de una forma completamente determinista: cambia de un estado interno a otro (quizá el mismo); reemplaza el 0 o el 1 que acaba de leer por el mismo o por un distinto símbolo 0 o 1; se mueve un cuadro hacia la derecha o la izquierda, finalmente, decide si continuar el cálculo o si terminarlo y detenerse. (Penrose, 1996, p. 41)

[10] “En realidad, en su descripción original Turing permitía que su cinta estuviera marcada de maneras más complicadas, pero esto no supone ninguna diferencia real. Las marcas complejas podrían descomponerse siempre en series de marcas y espacios en blanco. Me tomaré otras libertades sin importancia con respecto a las especificaciones originales de Turing” (Penrose, 1996, p. 40).

> Por ejemplo, para el algoritmo de Euclides nuestro dispositivo necesitará actuar sobre el *par* de números A y B. Se pueden diseñar, sin gran dificultad, máquinas de Turing que ejecuten este algoritmo. Como ejercicio, los lectores aplicados pueden verificar que la siguiente descripción de una máquina de Turing (que llamaré EUC) ejecuta realmente el algoritmo de Euclides cuando se aplica a un par de números unarios separados por un 0:
>
> 00 =>00D, 01=>11I, 10=>101D, 11=>11I,
>
> 100=>10100D, 101=>110D, 110 =>1000D, 111=>111 D,
>
> 1000=>1000D, 1001 => 1010D, 1010=>1110I,
>
> 1011=>1111I, 1100=>1100I, 1101 =>11I,
>
> 1110=>1110I, 1111=>10001I, 10000 =>10010I,
>
> 10001=>10001I, 10010=>100D, 10011 =>11I,
>
> 10100=>00ALTO, 10101=>10101D (p. 44)
>
> En el caso de EUC, para tener una idea más clara de lo que supone, puede ensayarse con algún par explícito de números, digamos 6 y 8. El dispositivo de lectura se encuentra, como antes, en el estado 0 e inicialmente a la izquierda, y la cinta estará ahora marcada inicialmente de la forma:
>
> ...000000000001111110111111100000...
>
> Cuando, después de muchos pasos, la máquina de Turing se detenga, tendremos una cinta marcada:
>
> ...000011000000000000... (Penrose, 1996, p. 45)

Según Penrose, las distintas operaciones aritméticas básicas pueden ser realizadas por la máquina de Turing, así mismo, pueden construirse máquinas de Turing sin una instrucción específica sobre la operación matemática que debe resolver, ya que dicha instrucción debe estar en la cinta. En pocas palabras, la máquina de Turing realiza operaciones mecánicas que arrojan resultados eficientes, lo que pone en cuestionamiento la inferioridad sobre el ser el humano.

1.2. Conexionismo

El modelo neuronal o conexionista aparece alrededor de la década de los años 80. Este modelo pretende alcanzar una mayor cercanía a la estructura del cerebro, a diferencia del modelo simbólico que centra su atención en la sintaxis y el leguaje como punto de partida.

> En los modelos conexionistas los contenidos mentales no se codifican ya en fórmulas sintácticas, sino en redes de actividad. Por tanto, si nuestra mente fuese una red conexionista, entonces parece que el modo como se instancian los contenidos mentales no podría ser, como propone la Imagen Sintáctica, cerebro - sintaxis => semántica, sino redes cerebrales - realizan redes conexionistas redes conexionistas = codifican => contenidos mentales (Corbí y Prades, 2007, p. 162)

Como ya se vio en líneas anteriores, el modelo simbólico intenta simular las capacidades cognitivas y las conductas humanas, a través del procesamiento de fórmulas sintácticas, "pues se entiende que tales capacidades descansan necesariamente en un sistema de representación cuyos elementos básicos reflejan la estructura sintáctica del lenguaje" (Corbí y Prades, 2007, p. 151). El experimento de la máquina de Turing, es un ejemplo concreto de lo que busca el modelo simbólico, una simulación e imitación de la conducta humana, a través de toda una estructura sintáctica. Empero, este modelo va a presentar falencias, por lo que las propuestas del modelo conexionista van a ser atractivas para la IA.

Según Josep E. Corbí y Josep L. Prades, uno de los problemas de los modelos simbólicos consiste en suponer que existen principios que determinan el contenido mental, y que tales principios están asociados a cada fórmula sintáctica. No obstante, se han expuesto argumentos que refutan este principio, ya que "hay ciertos rasgos peculiares de los contenidos mentales (y, en general, de la semántica) que impiden tal correlación" (Corbí y Prades, 2007, p. 158).

Un segundo problema que se evidencia, es la incapacidad para explicar cómo funciona realmente la mente. Aunque este modelo puede imitar muchas conductas humanas, tales como las mencionadas en los experimentos anteriores y, además, puede producir una buena cantidad de pensamientos a partir de una información limitada, no es suficiente para comprender el modelo como opera la mente humana

Una tercera dificultad radica en el problema del marco. Como bien se sabe, los seres humanos tienen la capacidad para seleccionar y clasificar información, así, no les cuesta hacer relaciones y obtener respuestas y acciones coherentes, eficaces e inmediatas. Un ejemplo que aplica para ilustrar la situación, tiene que ver con la capacidad para distinguir que el color de una bebida no afecta su sabor. No obstante, el modelo simbólico no responde de forma satisfactoria a este tipo de situaciones, para ello, es necesario introducir toda una serie de información que pueda responder de manera coherente a la situación específica. Un trabajo dispendioso para las múltiples combinaciones y relaciones que puede realizar la mente humana. Para Corbí y Prades (2007) "la opción de dotar al ordenador con un esquema nos obliga a pagar el precio de una excesiva rigidez que no concuerda con la flexibilidad de nuestras reacciones ante situaciones inesperadas, distintas de las tipificadas en el esquema" (p. 159).

Por su parte, los modelos conexionistas están compuestos por unidades que pueden estar activadas o desactivadas, y que a su vez, pueden estar conectadas entre sí. Para estos investigadores, "el valor de esas conexiones y su carácter (excitatorio o inhibitorio) va cambiando a lo largo del proceso de entrenamiento de la red" (Corbí y Prades, 2007, p. 160), por lo que se podría hablar de una progresión. En pocas palabras, un modelo conexionista es una red de redes de unidades o pautas de conexión. Así, una forma de comprender cómo estos sistemas pueden codificar información, es suponiendo que cada unidad representa un estado del mundo. Además, en estos modelos, los contenidos mentales no se codifican a través de fórmulas sintácticas como sucede en el modelo simbólico, lo hacen en redes de actividad.

Ahora bien, si no le es posible integrar sistemáticamente toda la información que adquiere del medio, esta no se paraliza como lo expresan Corbí y Prades, es usual que las unidades que se activan en un momento determinado representen rasgos del mundo y no sean compatibles entre sí. El modelo conexionista busca soluciones a dicho problema, y lo que intenta es integrar la información para arrojar un dato coherente. Si dicha situación se presentase en el modelo simbólico, no sabría cómo resolver el problema, ya que este modelo debe estar programado con anterioridad y cualquier situación imprevista termina siendo un fracaso.

Así, una ventaja del modelo conexionista es que este "puede dotarse a las redes neurales de reglas de aprendizaje que las lleva a generalizar espontáneamente cuando se detecta una cierta correlación (negativa o positiva) entre la presencia de varios rasgos del mundo" (Corbí y Prades,

2007, p. 161). Además, una similitud entre las redes conexionistas y las habilidades humanas, recae sobre la degradación a la que pueden llegar, debido a los condicionamientos.

Empero, estos modelos no dan una respuesta exacta de cómo funciona el pensamiento, lo que afecta el hecho de poder explicar el papel causal de los contenidos mentales, lo que presenta una desventaja frente a los modelos simbólicos que intentan analizar la mente desde un punto de vista cognitivo.

CAPÍTULO II

La filosofía de la mente de John Searle

El propósito central de la filosofía de la mente de John Searle radica en clarificar conceptos como el de conciencia, intencionalidad, subjetividad y causación mental, los cuales comprometen los fenómenos mentales, tales como deseos, creencias, sentimientos, emociones, etc.

Históricamente dichos conceptos se han utilizado de manera equívoca; las diferentes posiciones acerca del problema de la mente se han justificado bajo supuestos que, en su mayoría, están cargados de errores y falencias conceptuales, generando así mayores problemas a la hora de comprender el funcionamiento de la mente.

En el presente capítulo se ofrecerá una breve explicación sobre el tan cuestionado problema mente-cuerpo, así como una solución al mismo por parte de Searle; teniendo en cuenta los procesos de micronivel y macronivel, posteriormente, se analizarán conceptos como el de conciencia e intencionalidad, ambos utilizados, según Searle, de manera errónea en la filosofía tradicional de la mente.

2.1. Solución al problema mente-cuerpo

El problema mente-cuerpo ha sido un tema de interés filosófico para los estudiosos de las teorías de la mente. Searle no es ajeno a este problema, y antes de ofrecer una solución al mismo, se hace necesario realizar un breve bosquejo sobre el origen del dualismo y sus implicaciones.

Rene Descartes, un influyente pensador de la filosofía moderna del siglo XVII, acogió la teoría dualista, en la que sostenía que el mundo se dividía en dos clases diferentes de sustancias o entidades de existencia autónoma: por un lado, existen las sustancias mentales, y por otro, sustancias extensas.

El planteamiento de este dualismo tiene su origen en la conocida máxima del pensador: "*cogito ergo sum"* (pienso, luego existo), la cual ha generado gran controversia acerca de la certeza que parece poseer. Para Descartes el *cogito* y el *sum* van a ser considerados como iguales, pero independientes, es decir, como ideas claras y distintas.

Bernard Williams introduce unos términos que explican mejor la referencia de ideas claras y distintas que introduce Descartes. Así, tanto *cogito* como *sum* son incorregibles, o sea, "si alguien cree que piensa, o de nuevo que existe, entonces tiene necesariamente una creencia verdadera" (Williams, 1996, p. 93). Además, las dos se autoverifican, por lo que si alguien asevera la proposición tiene que ser verdadera. No obstante, para Williams (1996):

> El hecho de que algún grupo destacado de proposiciones sobre la vida mental sea incorregible o evidente, no significa que todas las proposiciones lo sean, o que la incorregibilidad y la evidencia sean cualidades necesarias de lo que sabemos sobre lo mental. (p. 105)

Una vez Descartes está convencido de la existencia del *cogito,* afirma que a este deben pertenecer una gran variedad de operaciones mentales que él experimenta, ya lo decía en su segunda meditación: "¿Qué soy, pues? Una cosa que piensa. ¿Qué es una cosa que piensa? Es una cosa que duda, entiende, concibe, afirma, niega, quiere, no quiere y, también, imagina y siente. Ciertamente no es poco, si todo eso pertenece a mi naturaleza" (Descartes, 2007, p. 131). El mismo Descartes experimenta sensaciones que le permiten dar razones sobre la existencia del *cogito*. Finalmente es él quien duda, siente, y es consciente de los objetos como si provinieran de los sentidos, ya que oye ruidos, observa y puede tener sensaciones de calor o frío.

Hasta el momento Descartes está convencido de que piensa, "soy yo, yo existo, [repite] eso es cierto, pero ¿Durante cuánto tiempo? Todo el tiempo en el que piense, pues quizás pueda suceder que, si cesara de pensar, cesaría así mismo de existir" (Williams, 1996, p. 13). En efecto, en la segunda meditación deja claro que su esencia es ser una cosa pensante, y que, si dejara de pensar, dejaría de existir.

> Como por un lado tengo una idea clara y distinta de mí mismo en tanto que cosa que piensa y que no tiene extensión. Y, por otro lado tengo una distinta del cuerpo en tanto que cosa que tiene extensión y que no piensa, es cierto que este yo, es decir mi alma, que hace que yo sea lo que soy, es entera y verdaderamente distinta de mi cuerpo, y puede ser y existir sin él. (Williams, 1996, p. 132)

Descartes llegó a la conclusión de que, si podía pensar que dudaba de la existencia de otras cosas, tenía que existir, sin embargo, tanto la mente como el cuerpo son concebidos como distintos. En palabas de Searle (2006), el filósofo francés:

> Creía que una sustancia debía tener una esencia o un rasgo esencial que la hacía ser lo que era (por cierto, toda esa jerga sobre la sustancia y la esencia proviene de Aristóteles). La esencia de la mente es la conciencia o el "pensamiento", como él la denominó; y la esencia del cuerpo es el hecho de extenderse en tres dimensiones del espacio físico: la extensión. (p. 28)

La sustancia mental de la que hablaba este pensador de la modernidad, definía la existencia, de hecho, afirmaba que siempre el ser humano se encuentra consciente y en el caso de no ser así, dejaría de existir. Además, la mente tiene una independencia total del cerebro. Por eso cuestionamientos sobre las relaciones causales entre la mente y el cuerpo (¿Cómo pueden los procesos cerebrales producir fenómenos mentales? ¿Cómo puede el cerebro ser la causa de la mente?) son equívocos para la filosofía de Descartes.

Por su parte, la sustancia extensa tenía diferentes modos o modificaciones en la cual podía manifestarse. Para Descartes todos los cuerpos eran divisibles, lo que significaba que se podían descomponer en partes más pequeñas y llegar a la destrucción de este; a diferencia de la sustancia mental, la cual era indivisible y eterna. Así todos los seres humanos eran considerados como entidades compuestas por una mente y un cuerpo, por dos sustancias diferentes.

> Por un lado tengo una idea clara y distinta de mí mismo en tanto que cosa que piensa y que no tiene extensión y, por otro lado, tengo una idea distinta del cuerpo en tanto que cosa que tiene extensión y no piensa. (Williams, 1996, p. 356)

Aunque este pensador reconocía que la mente causaba sucesos en el cuerpo y que sucesos del cuerpo causaban sucesos en la mente, no comprendía exactamente su funcionamiento, hasta que finalmente dio con la hipótesis de que existía un punto de conexión entre la mente y el cuerpo, y este era la glándula pineal; un pequeño órgano situado en la base del cráneo.

> El filósofo advirtió que todos los elementos cerebrales situados en un lado tenían su réplica en el otro. Debido a la existencia de los hemisferios, la anatomía parece mostrarse en duplicado. Pero como todos nuestros sucesos mentales ocurren en forma unitaria, debe haber en el cerebro algún punto unificado donde confluyen las dos corrientes. El único órgano no duplicado que Descartes pudo encontrar dentro del cerebro fue la glándula pineal, por la cual supuso que esta debía ser el punto de contacto de lo mental y lo físico. (Searle, 2006, p. 50)

Según Williams, un ser humano que tiene la facultad de pensar, sentir y tener experiencias perceptivas, es puramente consciente dentro de la concepción cartesiana. No obstante, no se tendrían tales experiencias si no se tuviera un cuerpo, a lo que Descartes responde que esas experiencias no son más que percepciones de estados del cuerpo, trasmitidas al alma a través de la glándula pineal.

Ahora bien, ¿qué dificultades trajo para la filosofía el dualismo sustancial qué proponía Descartes? Muchos fueron los problemas que se desencadenaron tras dicho planteamiento, incluso, aún resuenan en las teorías contemporáneas de la mente, como se verá más adelante.

Para Searle, dicho dualismo impide, en primer lugar, hacer una exposición coherente de las relaciones causales.

> El inconveniente de esta concepción es que, visto lo que sabemos sobre el funcionamiento del mundo, cuesta tomarla en serio como hipótesis científica. Sabemos que en los seres humanos la conciencia no puede existir en manera alguna sin ciertos procesos físicos que se desenvuelven en el cerebro. (Searle, 2006, p. 63)

Sin embargo, Descartes se cuestionaba sobre cómo pensamientos y sensaciones como el dolor podrían ser causados por sucesos físicos y viceversa.

Alrededor de esta concepción dualista se han planteado una cantidad de problemas que intentan clarificar esa dicotomía entre mente y cuerpo, y que buscan dar una solución a las relaciones de los seres humanos con el resto del universo. Según las exploraciones filosóficas de Searle, uno de los mayores inconvenientes para abordar el problema en cuestión, ha sido el desconocimiento sobre el funcionamiento del cerebro, además, las teorías postuladas basadas en supuestos poco fiables.

> ¿Qué es exactamente la neurofisiología de la conciencia? ¿Por qué necesitamos dormir? ¿Por qué exactamente nos emborracha el alcohol? ¿Cómo exactamente se almacenan los recuerdos en el cerebro? […] Muchas de las afirmaciones sobre la mente hechas en varias disciplinas que van desde la psicología freudiana a la inteligencia artificial dependen de este tipo de ignorancia. (Searle, 1995, p. 18)

El problema se agudiza cuando no se acepta que la mente consciente funciona como cualquier otro fenómeno biológico, y se empieza a creer que esta es inmaterial y subjetiva, a

diferencia de un órgano como el cerebro, el cual es algo material y localizable en el espacio. A esto se le agrega la creencia popularizada sobre los problemas de los cuales se puede ocupar la ciencia: “fenómenos objetivamente observables”. Realmente, la concepción materialista termina por negar la existencia de los estados subjetivos conscientes mentales, como se verá en líneas posteriores. Ahora bien,

> La cuestión que inquieta a Searle, una vez ha explorado diferentes teorías de la mente, tienen que ver con esa negación de las características intrínsecamente mentales a fenómenos mentales, lo que ha dilatado el problema mente-cuerpo, y no ha permitido una solución eficaz, sin negar los rasgos de los fenómenos mentales, los cuales no han sido muy bien aceptados y recibidos por nuestra concepción científica del mundo como compuesto de cosas materiales. (Searle, 1995, p. 19).

El primer rasgo de la mente es la conciencia, que Searle define de manera general, como el hecho central de la existencia humana, sin la cual no sería posible aspectos propiamente humanos como el lenguaje, los sentimientos, las actitudes, etc. Dicho de otro modo, “la conciencia es el hecho central de la existencia específicamente humana, puesto que sin ella todos los demás aspectos específicamente humanos de nuestra existencia -lenguaje, amor, humor y así sucesivamente- serían imposibles” (Searle, 1995, p. 20).

Para este filósofo el rasgo fundamental de lo mental es la conciencia, la cual concibe como una cualidad biológica de los cerebros humanos y de ciertos animales y, como una parte del orden biológico natural como cualquier otro. Searle ofrecerá una definición de la conciencia, desde el sentido común, a saber, la conciencia entendida como el sentir y el advertir. Así la conciencia tiene niveles de mayor o menor intensidad.

> Cuando me despierto, después de dormir sin haber soñado, paso a estar consciente, un estado que continua tanto tiempo como estoy despierto. Cuando voy a dormir, o me ponen bajo una anestesia general, o muero, mis estados conscientes cesan. Si sueño mientras duermo, adquiero de nuevo conciencia, aunque las formas de conciencia en el sueño son, en general, de un nivel de intensidad mucho menor que la conciencia ordinaria mientras estamos despiertos. (Searle, 1996, p. 95)

Un segundo rasgo de la mente es lo que filósofos y psicólogos llaman *intencionalidad.* Que es *“*el rasgo mediante el cual nuestros estados mentales se dirigen a, o son sobre, o se

refieren a, o son de objetos y estados de cosas del mundo distintos de ellos mismos" (Searle, 1995, p. 20), basta aclarar que no solo se refiere a intenciones, sino también a creencias, deseos, temores, etc.

El tercer rasgo de la mente es la subjetividad; un rasgo que ha sido objeto de disputa entre diferentes filósofos de la mente, quienes se niegan a incorporarlo en una concepción científica y objetiva de la realidad. No obstante, para Searle es un rasgo eminentemente mental, así:

> La subjetividad está marcada por hechos tales como que yo puedo sentir mis dolores y tú no puedes. Yo veo el mundo desde mi punto de vista, tú lo ves desde tu punto de vista. Yo soy consciente de mí mismo y de mis estados mentales internos, como algo completamente distinto de los yoes y los estados mentales de otras personas. (Searle, 1995, p. 2)

Por último, y no menos importante, se tiene el problema de la causación mental, donde los fenómenos mentales son causados por procesos que ocurren en el cerebro como procesos de macronivel. Uno de los ejemplos más comunes que propone Searle tiene que ver con la sensación de dolor, el cual está causado por una serie de eventos que comienzan en las terminaciones nerviosas libres y terminan en el tálamo y otras regiones del cerebro. "De hecho, por lo que respecta a las sensaciones efectivas, los eventos que acaecen dentro del sistema nervioso central son suficientes para causar dolores" (Searle, 1995, p. 23), de esta manera, todos los pensamientos y sensaciones propios están causados por procesos que ocurren dentro del cerebro, aun cuando los materialistas niegan estos rasgos de lo mental. Para ellos no puede existir algo ontológicamente subjetivo, lo que implica para Searle negar algo tan esencial como la experiencia, ya que esta es ontológicamente subjetiva.

Todos estos son rasgos de la vida mental, el problema radica en aceptar que hacen parte de esta, por tener la creencia de que no pueden ser observables y, por tanto, deben ser desechados de cualquier explicación científica seria. No obstante, si se mira con detenimiento qué produce un dolor, se da cuenta que está causado por un proceso neurofisiológico:

> La sensación afectiva de dolor parece estar causada tanto por la estimulación de las regiones basales del cerebro, especialmente el tálamo, y la estimulación del córtex somatosensorial. Nuestras sensaciones de dolor están causadas por una serie de eventos que comienzan en las terminaciones nerviosas libres y terminan en el tálamo y otras regiones del cerebro. (Searle, 1995, p. 23)

De acuerdo con esto, Searle reafirma que todos los pensamientos y sensaciones están causados por procesos que ocurren dentro del cerebro y, por lo tanto, los dolores son tan solo rasgos del cerebro. Si bien, lo expuesto anteriormente puede ser aceptado, el punto álgido de la discusión se produce tras el siguiente interrogante: "Cómo puede ser que el cerebro cause las mentes, y las mentes sean únicamente rasgos del cerebro?" (Searle, 1995, p. 24). Según Searle, uno de los inconvenientes que se genera frente al interrogante, es no tomar como verdadero el hecho de que el cerebro causa las mentes, y al mismo tiempo las mentes son únicamente rasgos del cerebro. Aunque esta causación suena a dualismo, no debe ser entendida de tal modo.

"Los fenómenos mentales, todos los fenómenos mentales, ya sean conscientes o inconscientes, visuales o auditivos, dolores, cosquilleos, picazones, pensamientos, toda nuestra vida mental, están efectivamente causados por procesos que acaecen en el cerebro" (Searle, 1995, p. 22). Los pensamientos y sensaciones, por ejemplo, están causados por procesos que ocurren en el cerebro. En pocas palabras, todos los fenómenos mentales son causados por el cerebro y realizados en él como procesos de macronivel.

Para ahondar en esa explicación causal entre la mente y el cerebro, Searle ofrece una explicación desde las teorías atómicas y evolucionistas, las cuales han sido centrales en la explicación científica del mundo. De acuerdo con la teoría atómica del universo, la mayoría de los fenómenos físicos son tan pequeños que podrían ser considerados como partículas, así elementos de mayor magnitud como un planeta, un asteroide o una mesa, están compuestos de partículas, y esas partículas de otras más pequeñas, "hasta que finalmente alcanzamos el nivel de las moléculas, compuestas de átomos, que, a su vez, se componen de partículas subatómicas. Son ejemplos de partículas, los electrones, los átomos de hidrógeno y las moléculas de agua" (Searle, 1996, p. 98). En este sentido, explicar la realidad desde las teorías atómicas,

> nos impone el requisito, de que muchos tipos de macrofenómenos sean explicables en términos de microfenómenos. Y esto, a su vez, tiene la consecuencia de que habrá niveles diferentes de explicación del mismo fenómeno, dependiendo de si vamos de derecha a izquierda —de lo macro a lo macro o de lo micro a lo micro— o de abajo arriba —de lo micro a lo macro. (Searle, 1996, p. 99)

De esta manera, Searle proporciona un modelo basado en la física (para explicar las relaciones causales entre la mente y el cerebro) en el que se evidencia la relación causal entre

estructuras de micro y macronivel. Así, una distinción común en física es aquella que se da entre estructuras de micro y macro propiedades de sistemas a pequeña y a gran escala. Por lo que muchas propiedades superficiales o globales pueden explicarse causalmente por las estructuras de micronivel.

Por ejemplo, la liquidez del agua: esta es causada por elementos de micronivel, o sea, una molécula compuesta por dos átomos de hidrógeno y uno de oxígeno. En resumidas palabras, la explicación a diferentes fenómenos se da causalmente por los elementos que hacen parte del micronivel. En este caso, la propiedad física de la liquidez del agua está causalmente definida por las partículas de micronivel. De igual forma,

> la solidez de la mesa que está delante de mí se explica por la estructura de enrejado ocupada por las moléculas de las que está compuesta. Similarmente, la liquidez del agua se explica por la naturaleza de las interacciones entre las moléculas de H_2O. Esos macrorasgos se explican causalmente por la conducta de elementos de micronivel. (Searle, 1995, p. 25)

Así como la liquidez del agua y la solidez de la mesa están causadas o son compuestas por procesos de micronivel, los fenómenos mentales, son causados por procesos de micronivel que tienen lugar en el cerebro. Las experiencias visuales, por ejemplo, son causadas por una serie de eventos de micronivel que suceden en el cerebro, como respuesta a la estimulación óptica externa del sistema visual. En el siguiente caso, la experiencia visual es causada un vasto número de neuronas:

> Una experiencia visual comienza con el "asalto de los fotones a las células fotorreceptoras de la retina, los familiares bastones y conos. Estas señales se procesan a través de al menos cinco tipos de células en la retina -células fotorreceptoras, horizontales, bipolares, amacrinas y ganglioneras-. Pasan a través del nervio óptico al núcleo genicular lateral, y de allí las señales son transmitidas al córtex estriado y posteriormente se difunden a través de las células extraordinariamente especializadas del resto del córtex visual, las células simples, las células complejas, las células hipercomplejas de al menos las tres zonas, 17 (la estriada), 18 (el área visual II), y (19 el área visual III). (Searle, 1992, p. 270)

Searle propone solucionar el problema mente-cuerpo a través de la relación causal entre procesos de micro y macronivel, además deja claro que la mente funciona como cualquier otro fenómeno biológico. En ese sentido, "las experiencias visuales y auditivas, las sensaciones táctiles,

el hambre, la sed y el deseo sexual son causados todos ellos por procesos cerebrales y tiene su realización en la estructura del cerebro, y todos ellos son fenómenos intencionales" (Searle, 1995, p. 23). Lo que significa que "los deseos, las creencias, los temores, etc., son fenómenos intencionales, causados por procesos cerebrales. Así mismo, debe existir alguien que los experimente, por lo que la subjetividad es un hecho objetivo de la biología, es necesario que alguien los experimente" (Searle, 1995, p. 30).

Teniendo en cuenta los argumentos expuestos, se advierte que, indudablemente Searle ofrece una explicación neurobiológica de los rasgos de lo mental. En este sentido, la tesis que va a proponer Searle es lo que él llama *naturalismo biológico*. Esta tesis sustenta una explicación científica de la conciencia, la cual es causada por procesos neurofisiológicos. De acuerdo con esto, la conciencia no es más que un rasgo de macronivel del cerebro que es causado por los rasgos de micronivel de este. "La conciencia existe en un nivel superior de neuronas, esto es, redes neuronales. La conciencia no existe en una neurona *x* o *y* simplemente, sino en grupos" (Morales, 2010, p. 51).

Para Searle, esa relación que se da entre los procesos de micronivel y macronivel, es una relación causal. Cuando la relación es micro-macro, como sucede, por ejemplo, con los procesos neurofisiológicos que causan la conciencia, existe una relación que Searle denomina *emergentismo.*

> Una definición rigurosa de lo que es el emergentismo (en filosofía de la mente) es la siguiente: al alcanzar cierto grado de complejidad, de las estructuras físicas surgen (o emergen) propiedades mentales. Estas no pueden, como en la superveniencia simple y llana, ser reducidas a sus condiciones de base. En este caso nos referimos a estructuras neurofisiológicas como estructuras de base. En su emergencia, las propiedades mentales tienen poder causal sobre aquello del nivel del que han emergido. (Morales, 2010, p. 51)

En la definición anterior se afirmaba que al alcanzar cierto grado de complejidad de las estructuras físicas surgen o emergen propiedades mentales. El cerebro y el sistema nervioso central, por ejemplo, hacen parte de un sistema de estructuras complejo; dichos sistemas tienen diversos niveles de propiedades (también pueden ser niveles de explicación metafísicos) los cuales pueden producir alteraciones cuando es de un nivel inferior a uno superior, dicha alteración se define como emergente. En palabras generales, el emergentismo sostiene que la mente ha emergido

de la evolución del cerebro, de tal manera que lo mental sería una propiedad del cerebro "que interactúa y está influenciada por ella, pero que, a su vez, posee estructura y leyes propias" (Ruiz Santos, 2011, p. 116).

2.2. La conciencia

En líneas anteriores se mencionaba que la conciencia es un rasgo de lo mental, para Searle, el rasgo fundamental, ya que los demás rasgos: intencionalidad, subjetividad y causación mental, solo pueden ser entendidos como completamente mentales a través de las relaciones con la conciencia.

Searle simplifica la conciencia en el sentir y el advertir, en otros términos, la conciencia como experiencia. Así mismo, la concibe como un atributo biológico de los cerebros humanos y de ciertos animales y, como una parte del orden biológico natural como cualquier otro, la cual se explica a partir de relaciones causales:

> La conciencia es una propiedad causalmente emergente de los sistemas. Es un rasgo emergente de ciertos sistemas de neuronas en el mismo sentido en que la solidez y la liquidez son rasgos emergentes de sistemas moleculares. La existencia de conciencia puede ser explicada por las interacciones causales entre elementos del cerebro al micronivel. (Searle, 1996, p. 122)

Para este pensador la conciencia se ubica dentro de una visión científica del mundo, donde las teorías atómica y evolucionista, ofrecen una explicación sobre el surgimiento de la conciencia. Según la teoría atómica de la materia, el universo está constituido de fenómenos físicos extremadamente pequeños (partículas) de modo que todo está formado por partículas, y estas a su vez de moléculas compuestas de átomos y partículas subatómicas.

> Son ejemplos de partículas, los electrones, los átomos de hidrógeno y las moléculas de agua. Como ilustran estos ejemplos, las partículas mayores están compuestas de partículas más pequeñas; y todavía hay una enorme incertidumbre y mucha discusión sobre la identificación de las partículas más pequeñas de todas. (Searle, 1996, p. 98)

Las partículas se organizan en sistemas mayores, en este sentido, los macrofenómenos son explicados en términos de microfenómenos, por lo que, además existen diferentes niveles de

explicación de un mismo fenómeno, dependiendo si va de lo micro a lo macro, de lo macro a lo macro, de lo micro a lo micro, etc. Una forma de ilustrar estos niveles (micro-macro),

> sería la de que el agua está hirviendo porque la energía cinética transmitida por la oxidación de los hidrocarbonos a las moléculas de H20 ha causado que se muevan tan rápidamente que la presión interna de los movimientos moleculares iguala la presión del aire exterior, la cual, a su vez, se explica por el movimiento de las moléculas de las que está compuesto el aire exterior. (Searle, 1996, p. 99)

Por lo que la conciencia puede ser explicada causalmente por las estructuras de micronivel, así los deseos, las creencias, los temores, etc., son fenómenos intencionales, causados por procesos cerebrales.

Además de ofrecer una explicación atómica de la conciencia, Searle añade los principios de la biología evolucionista. Durante años, ciertos tipos sistemas de seres vivos han evolucionado, dichos sistemas han estado compuestos de moléculas de carbono en las que abunda el hidrógeno, el nitrógeno y el oxígeno. Aunque la manera de evolucionar se ha tornado complicada, el proceso básico es que las instancias de los tipos de sistemas llevan a la existencia de otras instancias similares, así, cuando las instancias originales se destruyen, las otras instancias continúan, y así se repite el proceso una y otra vez. "Variaciones en los rasgos superficiales, o fenotipos, de las instancias les proporcionan mayores o menores posibilidades de supervivencia, en relación con los ambientes específicos en que se encuentran" (Searle, 1996, p. 10). De este modo, "la conciencia encuentra su lugar naturalmente como un rasgo fenotípico de ciertos tipos de organismos con sistemas nerviosos altamente desarrollados" (Searle, 1996, p. 102).

> Algunos de esos sistemas tienen vida, y esos tipos de sistemas vivos han evolucionado a lo largo de enormes periodos de tiempo. Entre ellos, algunos han desarrollado cerebros que son capaces de causar y mantener conciencia. La conciencia es, así, un rasgo biológico de ciertos organismos en exactamente el mismo sentido de "biológico" en el que la fotosíntesis, la mitosis, la digestión y la reproducción son rasgos biológicos de los organismos. (Searle, 1996, p. 105)

Hasta aquí Searle deja claro que la conciencia hace parte del orden biológico natural, y que además se considera un rasgo evolutivo como cualquier otro. Sin embargo, la conciencia queda por fuera de varias explicaciones que ofrecen algunos teóricos, porque según estos, el rasgo

subjetivo de la conciencia imposibilita una explicación científica de la misma. Esta es apenas una razón por la cual la niegan, pero hay otras como se verá en los siguientes párrafos.

Todos los estados mentales tienen un rasgo especial que no poseen otros fenómenos naturales: la subjetividad. Searle utiliza dicho término en un sentido ontológico y no epistemológico, así, todo estado consciente es siempre el estado consciente de alguien

> Del mismo modo en que tengo una relación especial con mis estados conscientes, que no es como mi relación con los estados conscientes de los demás, ellos tienen una relación con sus estados conscientes que no es igual a la relación que yo tengo con sus estados conscientes (Searle, 1996, pp. 106-107)

Al concebir el mundo formado de partículas y sistemas biológicos, de los cuales algunos de ellos son conscientes, se le otorga el carácter subjetivo a la consciencia. Searle da un ejemplo que permite comprender ese carácter ontológicamente subjetivo. "Tengo un dolor de espalda" este suceso es real y no depende de observadores, el suceso mismo tiene un modo de existencia subjetivo, en ese sentido, Searle afirma que la conciencia es subjetiva, y que este es un rasgo fundamental de la misma. Ahora bien,

> Supongamos que tratamos de reducir la sensación subjetiva consciente, de primera persona, de dolor a los patrones objetivos, de tercera persona, de actividad neuronal. Supongamos que intentáramos decir que el dolor no es "nada más que" los patrones de actividad neuronal. Si intentáramos tal reducción ontológica, los rasgos esenciales del dolor se dejarían de lado. Ninguna descripción de hechos objetivos, fisiológicos, de tercera persona, transmitiría el carácter subjetivo, de primera persona, del dolor, simplemente porque los rasgos de primera persona son diferentes de los rasgos de tercera persona. (Searle, 1996, p. 127)

Pero ¿Por qué la subjetividad se torna problemática para ofrecer una explicación científica de la conciencia? Según el filósofo en cuestión, explicar la subjetividad se ha tornado difícil por dos razones fundamentales: primero, la creencia equivocada de que hasta el último rasgo de la realidad ha de ser objetivo; segundo, la idea de que una realidad objetivamente observable presupone la noción de observación. No obstante, para Searle "la idea de que hay una observación de la realidad es precisamente la idea de representaciones (ontológicamente) subjetivas de la realidad. La ontología de la observación —como opuesta a su epistemología— es precisamente la

ontología de la subjetividad" (Searle, 1996, p. 110). En pocas palabras, la observación es siempre la observación de alguien, es la conciencia, la experiencia de lo que se observa. Pero las afirmaciones que se hacen sobre ellas pueden ser epistemológicamente objetivas

> Filósofos como Armstrong (1980) eliminan tácitamente la subjetividad al tratar la conciencia como una mera capacidad para realizar discriminaciones sobre los propios estados interiores y Changeux, el neurobiólogo francés, define la conciencia simplemente como "un sistema global regulatorio que trata de los objetos mentales y de las computaciones que usan esos objetos"[11] (Searle, 1996, p. 111)

Aunque ambos presuponen una concepción de tercera persona de la realidad, no se puede ofrecer únicamente una concepción epistemológicamente objetiva, sino también una ontológicamente subjetiva, donde haya lugar para la conciencia, y para una explicación de primera persona. De tal manera que "los procesos biológicos producen fenómenos mentales conscientes, y estos son irreductiblemente subjetivos" (Searle, 1996, pp. 109-110). Por su parte, Nagel presenta un argumento en el que señala que existe una limitación a la hora de concebir la subjetividad, pero que indudablemente la conciencia es subjetiva; así como existen relaciones entre fenómenos materiales con otros fenómenos materiales y dicha relación puede presentarse de manera subjetiva, también existen relaciones entre ciertos fenómenos materiales con fenómenos mentales, en los que un extremo de la relación ya es subjetivo.

Para Nagel (2000) "un organismo tiene estados mentales conscientes si y sólo si hay algo que es cómo es *ser* ese organismo, algo que es cómo es ser *para* ese organismo" (p. 46). Uno de los ejemplos más tradicionales de Nagel para explicar el carácter subjetivo de la experiencia es el del murciélago:

> Supongo que todos creemos que los murciélagos tienen experiencias. Después de todo, son mamíferos, y no dudamos que tenga experiencias, así como tampoco dudamos que los ratones, las palomas y las ballenas las tengan. […] los murciélagos tienen un rango de actividad y un aparato sensorial tan diferente del nuestro que el problema que deseo plantear resulta muy vívido. (Nagel, 2000, p. 49)

Aunque se conozca a profundidad la estructura fisiológica del murciélago, a la hora de imaginar cómo es *ser* un murciélago para un *murciégalo,* el ser humano se vería limitado a hacerlo,

[11] Changeux, J.P. (1985). *Neuronal Man: The Biology of Mind* (Trad. ing. L. Garey). New York: Pantheon Books.

tal como lo expresa Nagel, dado que las experiencias humanas no tienen ninguna relación con la de estos mamíferos. Del mismo modo que se puede saber cómo es ser un ciego o sordo de nacimiento, y esa persona tampoco puede saber la experiencia de una persona que escucha y ve normalmente. En pocas palabras el argumento de Nagel intenta mostrar que la subjetividad no se puede desligar de la experiencia, ya que esta es la que permite tener una vivencia de primera persona que otros no pueden experimentar de la misma manera. Por eso es imposible saber cómo es ser otro mamífero. Searle extrae el argumento de Nagel y argumenta lo siguiente:

> No podemos desembarazarnos de la subjetividad de nuestra conciencia, para ver su relación necesaria con su base material. Nos formamos una imagen de la necesidad basada en nuestra subjetividad, pero no podemos formarnos de ese modo una imagen de la necesidad de la relación entre la subjetividad y los fenómenos neurofisiológicos, porque ya estamos en la subjetividad, y la relación imaginativa requeriría que no saliéramos de ella. (Searle, 1996, p. 114)

Según Searle, la conciencia tiene como finalidad organizar un conjunto de relaciones entre el organismo, su entorno y sus estados internos. Dicha organización podría ser presentada por medio de las modalidades sensoriales, de modo que el organismo oye, ve, y obtiene información consciente de su entorno, "además de sus experiencias sensoriales conscientes, el organismo también tendrá experiencias características de actuar" (Searle, 1996, p. 118).

En resumidas palabras, se puede decir que en "la percepción consciente, el organismo tiene representaciones causadas por los estados de cosas del mundo y, en el caso de la acción intencional, el organismo causa los estados de cosas del mundo por medio de sus representaciones conscientes" (Searle, 1996, p. 118), por lo que la subjetividad juega un papel esencial que compromete los estados conscientes.

Aunque Searle concibe la conciencia como un rasgo biológico e irreductible, los filósofos tradicionales de la mente se niegan a aceptar la conciencia de este modo, en realidad, hablar de conciencia ocasionó problemas para filósofos y psicólogos en un pasado no muy lejano, como lo expresa el mismo Searle en su libro *El redescubrimiento de la mente*.

Pero ¿Cuál ha sido la razón por la que se ha tornado inconcebible la conciencia como un rasgo biológico como cualquier otro? Realmente han sido varias, en primer lugar, el desconocimiento e ignorancia sobre el funcionamiento del cerebro, los prejuicios filosóficos sobre

los que se han fundamentado diferentes teorías de la mente y, por último, y no menos importante, los conceptos tradicionales que siguen arraigados a las nuevas teorías. Todas estas razones han dejado a la conciencia en un completo misterio. En palabras de Searle (1996):

> El "misterio" de la conciencia hoy en día está en, aproximadamente, la misma situación en que estaba el misterio de la vida antes del desarrollo de la biología molecular o el misterio del electromagnetismo antes de las ecuaciones Clark-Maxwell. Parece misterioso porque no sabemos cómo funciona el sistema de la neurofisiología/conciencia, y un conocimiento de cómo lo hace eliminaría el misterio. (p. 113)

Diferentes teorías de la mente le atribuyen rasgos a la conciencia, que en su mayoría son equívocos para Searle, además hacen una reducción de la conciencia que el filósofo en cuestión no acepta. Para mencionar algunos errores de manera general, se mirará el caso del dualismo, el cual sostiene que la conciencia es irreductible porque en los términos que se fundamenta no puede ser tratada científicamente.

El dualismo sostiene que el mundo se divide en dos clases diferentes de sustancias o entidades de existencia autónoma: por un lado, existen las sustancias mentales, y por otro, sustancias extensas, lo que significa que la mente debe entenderse diferente al cuerpo. Aunque los dualistas convergen con Searle en cuanto a que la conciencia no es reductible, estos no tienen una explicación científica para la conciencia, por lo que han sido fuertemente criticados. Para Descartes, por ejemplo, la conciencia termina siendo autoconciencia y solo tiene autoconciencia quien tiene mente. De tal manera que, a la hora de tratar diferentes problemas, lo hacen de manera separada, en ese sentido, la conciencia termina siendo algo misterioso, de una naturaleza diferente a lo físico, algo incomprensible e inexplicable, y desde luego, contrario a una explicación científica y biológica de la conciencia como lo ha venido sosteniendo Searle.

El materialismo es una teoría que sostiene que la microestructura del mundo está conformada por partículas materiales. Aunque la mayoría de las teorías mentales, coinciden en que todo está formado por partículas materiales, ha existido un afán del materialismo por negar los rasgos de lo mental, dentro de ellos, la conciencia. Pero ¿Por qué no aceptar que los rasgos de lo mental hacen parte del orden biológico?

Searle acepta y reconoce que, si bien el mundo está formado por partículas físicas, hay hechos de la propia experiencia que no se pueden negar, uno de ellos es que la conciencia, y todos

los estados conscientes tienen propiedades fenomenológicas irreductibles. No obstante, uno de los problemas centrales de las teorías materialistas radica en la negación de la existencia de determinados rasgos mentales, inclusive, ven la conciencia como un problema, que en lugar de explicar terminan por negar. A continuación, se mencionarán algunas teorías materialistas en las que la conciencia se vuelve problemática.

El conductismo se presenta en dos variedades, por un lado, se tiene el conductismo metodológico y por el otro, el conductismo lógico:

> El conductismo metodológico es una estrategia de investigación en psicología, con la propuesta de que la ciencia psicológica debe consistir en el descubrimiento de las relaciones entre los *inputs* estimulativos y los *outputs* conductuales (Watson, 1925). Una ciencia empírica rigurosa, de acuerdo con este punto de vista, no hace referencia alguna a elementos introspectivos misteriosos o mentalistas. (Searle, 1996, p. 47)

Por su parte, el conductismo lógico sostiene que no existen elementos a los cuales referirse en la medida en que existe una forma de conducta, de esta manera "los fenómenos mentales en cuestión no consisten en que se den realmente ciertos fenómenos, sino, más bien, en ciertas disposiciones a la conducta" (Searle, 1996, p. 47). En otras palabras, el conductismo está asociado al funcionamiento del lenguaje mental o psicológico, el cual funciona en relación con un leguaje fisicalista, donde se describen hechos que son públicamente observables y describibles desde un punto de vista de la tercera persona.

> En el modo material de habla, el conductismo pretende que la mente es solo conducta y disposiciones para comportarse. En el modo formal, consiste en el punto de vista de que las oraciones sobre los fenómenos mentales pueden traducirse a oraciones sobre la conducta real o posible. (Searle, 1996, p. 47)

De tal forma que, si se afirma que la consecuencia está asociada a la experiencia, y esta es subjetiva, es decir, que el único que tiene acceso a su consciencia es el propio sujeto, no habría una traducción a un lenguaje fisicalista y público. El conductismo defiende que lo que no se puede traducir a un lenguaje fisicalista debe quedar por fuera de cualquier explicación científica seria; y ya se ha visto que un rasgo fundamental de la consciencia es la subjetividad.

Una de las objeciones que se le hacen al conductismo en general, es la negación de los fenómenos mentales, en la que se ve involucrada la experiencia subjetiva de pensar y sentir.

Además, solo existe una preocupación por la conducta observada. Otra objeción tiene que ver con la negación de las relaciones causales entre los estados mentales y la conducta

> Por ejemplo, al identificar el dolor con la disposición a la conducta de dolor, el conductismo deja de lado el hecho de que el dolor causa la conducta de dolor. Del mismo modo, si tratamos de analizar las creencias y los deseos en términos de conducta, ya no podemos decir que las creencias y los deseos causan la conducta. (Searle, 1996, p. 47)

La crítica que hace Searle al conductismo consiste en el hecho de negar la existencia de los estados mentales internos, lo que significa que va en contra de las experiencias propias de la condición humana. De acuerdo con esto, la conciencia termina siendo problemática para los conductistas, quienes se niegan a aceptar los rasgos mentales, y la conciencia es un rasgo fundamental de lo mental que involucra de forma total la experiencia.

Por su parte, la teoría de la identidad sostiene que los estados mentales son idénticos a los estados del cerebro y del sistema nervioso central, así, los dolores son idénticos a ciertos estados del sistema nervioso central. Searle concibe como incorrecto afirmar que los dolores sean necesariamente estados cerebrales, "es incluso posible concebir una situación en la que yo tuviera este mismo dolor sin tener este estado cerebral y en la que tuviera este estado cerebral sin tener dolor" (Searle, 1996, p. 53). Para los teóricos de la identidad, la conciencia puede ser reductible, sin embargo, esta se vuelve problemática y difícil de definir. Pero ¿Por qué la conciencia se vuelve problemática para los materialistas? El rasgo característico de la conciencia es la subjetividad, no obstante, los materialistas se niegan a aceptar tal rasgo:

> Los materialistas se muestran reluctantes a aceptar ese rasgo porque creen que aceptar la existencia de la conciencia subjetiva sería inconsistente con su concepción de cómo debe ser el mundo. Muchos creen que, dados los descubrimientos de las ciencias físicas, lo único que podemos aceptar es una concepción de la realidad que niegue la existencia de la subjetividad. De nuevo, como sucedía con "conciencia", una manera de enfrentarse a la situación es la de redefinir "subjetividad" de modo que ya no signifique subjetividad, sino algo objetivo. (Searle, 1996, p. 68)

En este sentido, la explicación de la conciencia quedaría por fuera para los teóricos de la identidad, ya que la experiencia no tiene lugar en las explicaciones neurofisiológicas según estos;

un dolor, por ejemplo, no es nada más que disparos en las fibras C, el rasgo subjetivo de lo mental no tiene valor alguno, ya que ese dolor solo son disparos en las fibras C y de tal modo es descrito.

Otras teorías materialistas como el funcionalismo y la inteligencia artificial fuerte convergen en la idea de que la mente es solo un programa de ordenador. "Para el funcionalismo, los procesos mentales son procesos internos con un rol causal sobre la conducta, que constituyen funciones mediadoras entre entradas sensoriales y salidas motoras" (Martínez-Freire, 1996, p. 96). De tal manera que un estado mental es reducible a una estructura funcional, dicha estructura está asociada causalmente entre *inputs* del ambiente que causan procesos internos en el sistema cognitivo, y dichos procesos causan *outputs* conductuales.

Para la mayoría de los funcionalistas los procesos internos para ser explicados empíricamente deben ser físicos, además, tal descripción solo se puede hacer desde un punto de vista de tercera persona. Esto permite concluir que una explicación de la conciencia queda por fuera de la concepción funcionalista, ya que dentro de las estructuras causales no hay lugar para la conciencia. El único interés de estos teóricos es que los *inputs* y los *outputs* estén conectados de manera correcta a través de un conjunto de procesos internos.

La inteligencia artificial fuerte sostiene que la mente es simplemente un programa de ordenador:

> Dado que un programa puede ser implementado en cualquier *hardware* (con la única condición de que el *hardware* sea lo suficientemente estable y poderoso como para ejecutar los pasos del programa), los aspectos específicamente mentales de la mente pueden ser especificados, estudiados y comprendidos sin saber cómo funciona el cerebro. (Searle, 1996, p. 5)

Una objeción que se le hace a la inteligencia artificial fuerte es que el modelo computacional de la mente dejaba de lado aspecto cruciales como la conciencia y la intencionalidad. Objeción que será ampliamente expuesta en el siguiente capítulo.

2.3. Intencionalidad

Otro rasgo de lo mental es la intencionalidad, esta es "el rasgo mediante el cual nuestros estados mentales se dirigen a, o son sobre, o se refieren a, o son de objetos y estados de cosas del mundo distintos de ellos mismos" (Searle, 1995, p. 20). Según Searle, los estados Intencionales

representan[12] objetos y estados de cosas del mundo, del mismo modo, que los actos de habla representan objetos y estados de cosas del mundo. "Así, como mi enunciado de que está lloviendo es una representación de cierto estado de cosas, mi creencia de que está lloviendo es también una representación del mismo estado de cosas" (Searle, 1992, p. 26). Por lo que cada estado intencional consta de un contenido representativo, en otros términos, que cuando se afirma que una creencia es una representación, se está diciendo que esta tiene un contenido proposicional, el cual representa un conjunto de condiciones de satisfacción.

Las condiciones de satisfacción se aplican a los estados Intencionales en los que hay una dirección de ajuste. Así, las creencias[13] y los enunciados pueden ser verdaderos o falsos dependiendo de si se cumplen o no. "*Creo que hoy va a llover",* esta creencia será verdadera, si y solo si, llueve, por lo que la dirección de ajuste en este caso es de mundo a mente; por su parte, los deseos e intenciones no pueden ser ni verdaderos ni falsos, simplemente se cumplen o se satisfacen, por lo que su dirección de ajuste, en este caso, es de mundo a mente. "*Le ordeno que por favor cierre la puerta"*, esta orden puede ser o no ser cumplida.

> Además hay también estados Intencionales que tienen la dirección de ajuste nula. Si yo siento haberte insultado o me alegro de que ganaras el premio, entonces, aunque mi pensar contiene una creencia de que te insulté y un deseo de no haberte insultado y mi alegría tiene una creencia de que tú ganaste el premio y un deseo de que ganaras el premio, mi pensar y mi alegría no pueden ser verdaderos o falsos del modo en que pueden serlo mis creencias, ni ser satisfechos del modo en que pueden serlo mis deseos. Mi pensar y mi alegría pueden ser apropiados o inapropiados. (Searle, 1992, p. 24)

De acuerdo con lo anterior, se puede decir que una creencia se satisface, si y solo si, es verdadera, al igual que una orden es satisfecha, si y solo si, es obedecida. Así como los estados Intencionales con contenido proposicional y una dirección de ajuste representan diversas condiciones de satisfacción, los actos de habla con un contenido proposicional y una dirección de

[12] "El sentido de 'representación' en cuestión pretende ser enteramente agotado por la analogía con los actos de habla: el sentido de 'representar' en el que una creencia representa sus condiciones de satisfacción es el mismo sentido en el que un enunciado representa sus condiciones de satisfacción" (Searle, 1992, p. 27).

[13] "Una creencia es un contenido proposicional en un cierto modo psicológico, su modo determina una dirección de ajuste mente -a- mundo, y su contenido proposicional determina un conjunto de condiciones de satisfacción. Los estados intencionales tienen que ser caracterizados en términos intencionales si no queremos perder de vista su intencionalidad intrínseca" (Searle, 1992, p. 27) Cabe aclarar que la Intencionalidad intrínseca hace referencia a la representacionalidad de los estados mentales que no depende de algo más para poder existir, es decir, que existen por sí mismo y que son propios de un ser pensante y racional, tales como creencias, deseos, etc.

ajuste también lo hacen, es decir, los actos de habla con un contenido proposicional están dirigidos a algo.

> Lo que es de crucial importancia ver aquí es que para cada acto que tiene una dirección de ajuste el acto de habla se satisfará si y solo si el estado psicológico expresado se satisface, y las condiciones de satisfacción del acto de habla y el estado psicológico expresado son idénticas. (Searle, 1992, p. 26)

Ahora bien, es importante tener en cuenta que los estados Intencionales determinan sus condiciones de satisfacción dada una Red de otros estados Intencionales respecto a un Trasfondo. Para que exista una creencia o un deseo, debe existir una gran cantidad de creencias y deseos que constituyen la Red de la creencia o el deseo en cuestión. Searle plantea el siguiente ejemplo para explicar cómo los estados Intencionales tienen un Trasfondo, o sea, estados preintencionales.

> Supongamos que hubo un momento particular en el que Jimmy Carter, primero, se formó el deseo de presentarse como candidato a la presidencia de los Estados Unidos, y supongamos, además, que ese estado intencional se realizó de acuerdo con las teorías de la ontología de lo mental preferidas por todo el mundo: él se dijo: "quiero ser candidato a la presidencia de los Estados Unidos"; tuvo cierta configuración neuronal en una cierta parte de su cerebro que produjo su deseo, lo pensó sin palabras e, impávido, tomó la resolución siguiente: "quiero hacerlo", etc. Ahora supongamos además que exactamente esas realizaciones de ese mismo estado mental, idénticas a las del caso anterior por lo que respecta a su tipo, ocurren en la mente y el cerebro de un hombre Pleistoceno que vive en una sociedad de cazadores – recolectores que vive en una sociedad desde hace miles de años. Este hombre tuvo una configuración neuronal idéntica a lo que corresponde al deseo de Carter, se encontró emitiendo la secuencia fonética: "Quiero ser candidato a la presidencia de los Estados Unidos", etc. Todo igual; sin embargo, por muy idénticas que sean respecto a su tipo las dos realizaciones el estado mental del hombre Pleistoceno no podría haber sido el deseo de ser candidato a la presidencia de los Estados Unidos. (Searle, 1992, p. 35)

Si bien cada hombre tuvo la misma "configuración neuronal" de querer ser presidente de los Estados Unidos, solo Carter tuvo el deseo, el cual debe estar inmerso en una Red de otros estados Intencionales. Dicho de otro modo, para que existiera tal deseo, debió haber tenido una serie de creencias sobre las formas de funcionamiento de la presidencia, las elecciones de ese país,

las condiciones para ser presidente, etc. en pocas palabras, debe tener unos saberes prácticos y teóricos que le permitan formarse tal deseo. Por su parte, el hombre del Pleistoceno no pudo haber tenido esas habilidades de Trasfondo, porque no tiene un saber ni práctico ni teórico sobre las condiciones que implican asumir la presidencia de los Estados Unidos.

Lo que resalta Searle del anterior ejemplo, es que los estados Intencionales son parte de una Red de estados Intencionales y solo tienen sus condiciones de satisfacción en relación con esa Red. Un deseo como el de querer ir a X lugar está configurado por otra serie de creencias que configuran o permiten tal deseo, esas creencias son las que conforman la Red y configuran la Intención de ir a X lugar. “Además de la Red de representaciones, hay también un Trasfondo de capacidades mentales no-representacionales; y, en general, las representaciones sólo funcionan, sólo tienen las condiciones de satisfacción que tienen, respeto de ese trasfondo no representacional” (Searle, 1992, p. 35).

Searle relaciona la tesis del Trasfondo con la Red y admite que los estados Intencionales requieren para su funcionamiento una Red de estados Intencionales. Pero ¿Cuál es la diferencia entre el Trasfondo y la Red? Para Searle esta pregunta termina siendo ociosa, y admite que una línea divisoria entre la Red y el Trasfondo no existe, porque la Red es aquella parte del Trasfondo que causa intencionalidad. No obstante, hay que tener presente que la Red tiene una parte no intencional, por lo que una diferencia entre la Red y el Trasfondo es que no toda la Red funciona intencionalmente. Así, hay un aspecto de la Red que funciona como condición de posibilidad y es preintencional.

Ahora bien, los estados Intencionales con una dirección de ajuste tienen contenidos que determinan sus condiciones de satisfacción, dicho de otra manera, tales estados tienen un contenido y determinan sus condiciones de satisfacción en relación con otros estados Intencionales. Por lo que estados mentales como creencias y deseos están configurados por otros estados como esperanzas, temores, sentimientos de frustración etc. En palabras de Searle, una Red holística, en el sentido de que un individuo tiene muchos estados mentales. Es imposible no aludir a esos estados Intencionales sin tener en cuenta esa Red, las razones que expone el filósofo son las siguientes:

> En primer lugar, porque una gran parte, quizá la mayor, de la Red está sumergida en el inconsciente y nosotros no sabemos del todo cómo sacarla a flote. En segundo lugar, porque

> los estados de la Red no se individualizan; no sabemos, por ejemplo, cómo contar creencias. Pero, en tercer lugar, si encontraríamos a nosotros mismos formulado una serie de proposiciones que parecerían "sospechosas" porque son, en algún sentido, demasiado fundamentales para ser calificadas como creencias, e incluso como creencias inconscientes. (Searle, 1992, p. 151)

El Trasfondo termina siendo entonces un conjunto de capacidades mentales no representacionales que dan lugar a las representaciones. En este sentido, los estados Intencionales tienen las condiciones de satisfacción que tienen, debido a un Trasfondo.

Además, el Trasfondo proporciona un conjunto de condiciones capacitadoras que hacen posible que funcionen formas particulares de Intencionalidad. Searle afirma que de la misma manera que la Constitución de Estados Unidos capacita a ciertas personas para formarse la creencia de querer aspirar a la presidencia, de igual forma, el Trasfondo capacita para tener unas formas de Intencionalidad. "En términos tradicionales, el Trasfondo proporciona las condiciones necesarias, pero no suficientes, para comprender, creer, desear, intentar, etc., y en este sentido es capacitador y no decisivo" (Searle, 1992, p. 166).

Entre los argumentos que expone Searle a favor del Trasfondo, se encuentra el que demuestra como los estados Intencionales necesitan de este. Para este pensador, una misma expresión puede ser interpretada de manera diferente, porque cada una se interpreta teniendo en cuenta un Trasfondo de capacidades humanas, las cuales contribuyen a diferentes interpretaciones, es decir, que el significado de una palabra, por ejemplo, está determinado por un contexto en el que existen un conjunto de habilidades, capacidades, disposiciones y potencialidades que permiten configurar los estados mentales.

Uno de los argumentos que expone Searle a favor de la hipótesis del Trasfondo tiene que ver con el significado literal, ya que todas las oraciones, desde la más sencilla hasta la más compleja, requieren de un Trasfondo preintencional, el cual permite interpretar de diversas maneras una misma expresión. Se aludirá a uno de los ejemplos que plantea Searle en su libro *El redescubrimiento de la mente*, con el fin de tener una mayor claridad sobre el argumento a favor de la hipótesis del Trasfondo:

> "Sam corta la hierba", "Sally corta el pastel", "Bill corta la tela", "Él corta su piel", se verá que la palabra "corta" significa lo mismo en cada una de ellas. […] Si digo "Sally corta con

> Bill", "La CNN corta sus emisiones mañana", "El rector corta la calefacción debido al plan de austeridad", en cada uno de los casos la palabra "corta" tiene un uso no literal. (Searle, 1996, p. 184)

Ahora bien, aunque en cada expresión se haga un uso literal del verbo "cortar", en cada expresión se entiende de manera diferente, dado un conjunto de capacidades y habilidades. Si se ordena que corte la tela, y que además se corte con alguien, por ejemplo, se sabe cómo hacerlo; y se sabe que en cada caso la palabra cortar funciona de manera diferente. dicho saber se debe a un Trasfondo local, el cual está determinado culturalmente. En pocas palabras, la expresión "cortar" será interpretada de manera diferente en las diferentes oraciones, porque cada oración se interpreta teniendo en cuenta un conjunto de capacidades y habilidades humanas para ciertas prácticas; en palabras de Searle "un saber-cómo, modos de hacer las cosas" (Searle, 1996, p. 185), y ese saber-cómo es el que contribuye en la comprensión de las palabras cortar la tela y cortar con alguien.

Otro argumento a favor de la hipótesis de Trasfondo tiene que ver con las oraciones que gramaticalmente están perfectamente construidas, y aunque se comprenden los significados de las palabras que la conforman, no se comprende la oración. "Así, por ejemplo, si uno oye la oración 'Sally corta la montaña', 'Bill corta el sol', 'Joe corta el lago', o 'Sam corta el edificio', se encontrará perplejo respecto de lo que esas oraciones, pueden significar" (Searle, 1996, p. 186), y no es comprensible porque se tiene un conjunto de capacidades y un saber-cómo, que imposibilita el si quiera pensar cómo cortar el sol. Searle afirma que, en este caso, se tendría que inventar una práctica de Trasfondo que fijase una interpretación en cada una de las oraciones. Para resumir el argumento del significado literal, Searle expresa lo siguiente:

> El significado literal (dejando de lado la indexicalidad y otros rasgos dependientes del contexto) determina las condiciones de verdad absolutamente y de modo aislado. Pero los significados literales son vagos, y las descripciones literales son siempre incompletas. Se añade una mayor comprensión y completitud complementando el significado literal con suposiciones y expectativas colaterales. Así, por ejemplo, cortar es cortar lo hagas como lo hagas, pero esperamos que la hierba se corte de una manera y los pasteles de otra. Así, si alguien dice: "Ve y corta la montaña", la respuesta correcta no es "No lo entiendo". ¡Naturalmente que entiendes esa oración castellana! Más bien la respuesta correcta es "¿Cómo quieres que la corte?". (Searle, 1996, p. 188)

CAPÍTULO III

La formulación del argumento de la habitación china de John Searle

Una de las preguntas fundamentales en IA reposa sobre la posibilidad de que las máquinas piensen, de hecho, quienes se han dedicado al estudio de la IA han tenido presente esta pregunta: "¿Pueden las máquinas pensar?" Aunque esta pregunta ha generado una discusión de carácter filosófico, Searle considera que se convierte en algo trivial, y una de las razones que expone, tiene que ver con el hecho de que una máquina es un sistema físico capaz de realizar ciertas operaciones, y eso no está muy alejado de visionar al ser humano como máquinas "podemos interpretar la materia que tenemos dentro de nuestras cabezas como una máquina de carne" (Searle, 1995, p. 41), en ese sentido, "todos los seres humanos son máquinas" que pueden pensar. Descartes ha vislumbrado el comportamiento de dichos organismos y el de los demás animales como procesos maquinales, en la medida que la mayoría de estos no son controlables por la propia voluntad, sino que se desarrollan de manera puramente automática:

> Supongo que el cuerpo no es otra cosa que una estatua o máquina de tierra a la que Dios da forma con el expreso propósito de que sea lo más semejante a nosotros, de modo que no solo confiere a la misma el color en su exterior y la forma de todos nuestros miembros, sino que también dispone en su interior todas las piezas requeridas para lograr que se mueva, como, respire y, en resumen, imite todas las funciones que nos son propias, así como cuentas podemos imaginas que no provienen sino de la materia y que no dependen sino de la disposición de los órganos. (Descartes, 1980, p. 120)

Searle considera que la formulación de la pregunta debería ser diferente: "¿Podría una máquina hecha por el hombre pensar?" (Searle, 1995, p. 4). Si se mira la naturaleza de esta pregunta, también termina siendo trivial, dado que es posible la clonación de seres vivos, en ese sentido es posible construir una máquina biológica igual a cualquier ser humano con los mismos poderes causales de producción de estados mentales.

Otra formulación de la pregunta sería:

> ¿Puede pensar un computador digital? […] Desde un punto de vista matemático cualquier cosa puede describirse *como si* fuera un computador digital. Y esto es así porque puede

> describirse como si instanciase o llevase o llevase a cabo un programa de computador. (Searle, 1995, p. 42)

Pese a que, "cualquier cosa es un computador digital, ya que cualquier cosa puede describirse como si llevase a cabo un programa de computador" (Searle, 1995, p. 42). Por lo tanto, la pregunta también resulta trivial, ya que es válido afirmar que los cerebros son computadoras digitales, en el sentido de que pueden llevar a cabo un número cualquiera de programas de computador, y desde luego los cerebros pueden pensar. Por lo que la pregunta que se debería plantear es: "¿Puede un computador digital, tal como se ha definido pensar?" Es decir, "¿es suficiente para, o constitutivo de, pensar el instanciar o llevar a cabo el programa correcto con los *inputs* y *outputs* correctos?" (Searle, 1995, p. 42).

Si bien un computador digital se va a definir como un *hardware* que es capaz de instanciar un programa; el punto clave de la discusión radica en el programa correcto con los *inputs* y los *outputs* adecuados, ya que el programa supone lo que sería la mente. Para Searle entonces, la respuesta a la anterior pregunta es negativa, y la razón que expone tiene que ver con la incapacidad de generar significados. El programa está definido de manera puramente sintáctica, y el pensamiento no consiste únicamente en manipular símbolos, sino también en atribuir significados, por lo que resulta pretensioso afirmar que un programa puede pensar.

Además, aunque se simule cualquier proceso mental, no se podría afirmar por ello, que el programa piensa en un sentido literal, en efecto, Searle plantea una serie de ejemplos donde las simulaciones son solo simulaciones y no algo real, de tal manera que un programa que simule procesos mentales, no piensa en un sentido literal como se defiende en IA fuerte.

> Podemos hacer una simulación computacional de las tormentas en los términos municipales del país, o de los incendios de los almacenes de Madrid. Ahora bien, en cada uno de esos casos, nadie supone que la simulación computacional es efectivamente la cosa real; nadie supone que una simulación computacional de una tormenta nos deje a todos mojados, o que sea probable que una simulación computacional de un incendio vaya a quemar una casa. (Searle, 1995, pp. 43-44)

Ahora bien, ¿qué se entiende realmente por IA? Searle, por su parte, ofrece una respuesta a este interrogante. Afirma que el funcionamiento del cerebro se compara con el funcionamiento de una computadora digital, por lo que la mente termina siendo un programa de computador; la IA

en sentido fuerte determina que “la mente es al cerebro lo que el programa es al *hardware* del computador” (Searle, 1995, p. 33). En este sentido, el mismo autor sustenta: “El cerebro es un número indefinidamente extenso de géneros de *hardware* de computador” (p. 34), que según la definición de IA podrían servir de base a los programas que constituyen la inteligencia humana.

Esta definición de IA artificial conlleva a una primera conclusión: “Cualquier sistema físico que tuviese el programa correcto con los *inputs* y los *outputs* correctos tendría una mente en exactamente el mismo sentido en que tú y yo tenemos mentes” (Searle, 1995, p. 34), en pocas palabras, una máquina que tuviera los *inputs* y los *outputs* adecuados podría pensar, por lo que la solución a la pregunta inicial acerca de si las máquinas pueden pensar, tiene que ver con la adecuada utilización del programa.

Pensadores como Herbert Simon y Alan Newel afirman que ya existen máquinas que literalmente pueden pensar, en el mismo sentido que los seres humanos lo pueden hacer. Para Newel, por ejemplo,

> La inteligencia artificial es un asunto de manipulación de símbolos físicos; no tiene ninguna conexión esencial con ningún género especifico de *wetware* o *hardware* biológico o físico. Más bien cualquier sistema que sea capaz de manipular símbolos físicos de una manera correcta es capaz de inteligencia en el mismo sentido literal que la inteligencia humana de los seres humanos. (Searle, 1995, p. 35)

Está claro que para estos teóricos la manipulación de símbolos constituye el rasgo esencial de la inteligencia, y en ese sentido, cualquier máquina que sea capaz de manipular símbolos tiene inteligencia. MacCarthy, por su parte, afirma que las máquinas tienen creencias, y que toda máquina que sea capaz de resolver problemas puede tener creencias; incluso, asevera que el termostato tiene creencias, convencido de que es capaz de creer en qué momento la temperatura cambia: “hace mucho calor acá”.

Las anteriores aseveraciones en defensa de la IA llevan a una segunda conclusión: “No hay nada esencialmente biológico en lo que respecta a la mente humana” (Searle, 1995, p. 35), ya que no puede ser tratada como un producto concreto de procesos biológicos de la misma manera que otros productos biológicos, según Searle. En pocas palabras, el cerebro termina siendo un *hardware* entre otros *hardware* de computador, como lo expresa el filósofo en su libro *Mentes, cerebros y ciencia.*

3.1. Distinción entre inteligencia artificial “fuerte” y “débil”

Aunque la definición de IA determina que la mente es como un programa de computador, existe una diferencia entre lo que se denomina IA fuerte e IA débil.

La IA fuerte concibe la mente como algo puramente formal, en este sentido,

> La computadora no es tan sólo una herramienta para estudiar la mente; más bien la computadora programada adecuadamente es realmente una mente en el sentido de que puede decirse literalmente que las computadoras que cuentan con programas correctos *comprenden y* poseen otros estados cognoscitivos. En la IA fuerte, como la computadora programada cuenta con estados cognoscitivos, los programas no son meras herramientas que nos permiten probar las explicaciones psicológicas, sino que los programas constituyen por si mismos las explicaciones. (Searle, 1994, p. 82)

Por su parte, la IA débil afirma que los computadores pueden simular los procesos cognoscitivos de los seres humanos. En este sentido, una computadora digital puede simular estados conscientes tales como creencias, deseos, etc. Según la definición que ofrece Searle en su libro *Mentes, cerebros y programas* “el principal valor que tiene la computadora en el estudio de la mente es que nos proporciona una herramienta muy poderosa que nos permite, por ejemplo, formular y comprobar hipótesis de un modo más riguroso y preciso” (Searle, 1994, p. 82).

3.2. Argumento de la habitación china

Grosso modo, el argumento de la habitación china de Searle muestra esencialmente la incapacidad que tiene un programa instanciado por una máquina para comprender. Antes de ampliar esta aseveración, se verá una primera formulación del argumento de la habitación china que expuso Searle en *Mentes, cerebros y programas:*

> Supongamos que estoy encerrado en una habitación y que se me proporciona un fajo grande de textos escritos en chino. Supongamos además (como es de hecho el caso) que no sé chino, ni escrito ni hablado, y que ni siquiera tengo la certeza de poder reconocer la escritura china como tal, distinguiéndola, por ejemplo, de la escritura japonesa o de una serie de garabatos sin significado alguno. Para mí, la escritura china solo es una serie de garabatos sin sentido. Ahora bien, supongamos que después de recibir este fajo de textos en chino se me proporciona otro que tiene una serie de reglas para relacionar el segundo fajo con el primero. Las reglas están en inglés y las entendiendo tan bien como cualquier otro hablante

> de esta lengua. Me permiten establecer una correlación entre un conjunto de símbolos formales y otro conjunto de símbolos formales. […] Supongamos también ahora que recibo un tercer fajo de símbolos chinos con algunas instrucciones, otra vez en inglés, que me permiten correlacionar elementos de este tercer fajo con los dos primeros, y que estas reglas me instruyen acerca de cómo responder con ciertos símbolos chinos de cierta forma a ciertos tipos de forma que me fueron proporcionadas en el tercer fajo. […] Imaginemos que estas personas también me proporcionan relatos en inglés, que yo comprendo, y luego me hacen preguntas en inglés sobre esos relatos y yo les respondo también en inglés. Supongamos también que después de un rato logro seguir también las instrucciones para descifrar los símbolos chinos y los programadores logran escribir tan bien los programas desde un punto de vista externo -esto es, desde la perspectiva de alguien que se encuentra fuera de la habitación donde estoy encerrado- que mis respuestas a las preguntas no pueden distinguirse en absoluto de las que proporcionaría un hablante chino. (Searle, 1994, p. 84)

Desde el punto de vista de la IA fuerte la computadora programada puede comprender relatos, no obstante, esta afirmación se queda sin sustento según Searle, ya que, si bien puede ofrecer respuestas satisfactorias en chino sin ser un hablante nativo, no puede tener una comprensión como tal de dichas respuestas. “Mis entradas y salidas de información son indistinguibles de las de un hablante chino y aunque pueda contar con el programa formal que se requiera, sigo sin entender nada” (Searle, 1994, p. 85).

Otra característica que se le atribuye a la IA fuerte es que el programa explica la comprensión humana, y, sin embargo, es posible ver cómo la computadora y su programa solo permiten una “traducción formal” de símbolos, pero sigue sin haber una compresión como tal. Ahora bien “¿Proporciona siquiera una condición necesaria o hace una contribución significante para la comprensión?” (Searle, 1994, p. 85), los defensores de la IA fuerte argumentan que cuando se comprende un relato en la lengua nativa, lo que se hace es manejar símbolos, de la misma forma que se manipulan los símbolos chinos para ofrecer respuestas como en el anterior experimento. “Es tan sólo manipulación simbólica más formal lo que distingue el caso del inglés, que sí comprendo, del chino, que no comprendo” (Searle, 1994, p. 85), según Searle, la credibilidad de este argumento tiene que ver con la suposición de crear un programa que tenga las mismas entradas y salidas de los hablantes nativos. Aun cuando, mientras que los programas se definan en términos de cómputo y manipulación de símbolos, estos por sí solos no permiten la comprensión. Así,

“cualesquiera que sean los principios puramente formales que se introducen en la computadora, no serán suficientes para la comprensión, ya que un ser humano puede seguir los principios formales sin entender nada” (Searle, 1994, p. 86).

Algunos defensores de la IA no están de acuerdo con los argumentos que expone Searle en el experimento de la habitación china, para mostrar que las computadoras carecen de un rasgo fundamentalmente humano como lo es la comprensión, en consecuencia, ofrecen una serie de réplicas que el mismo Searle intenta refutar como se verá a continuación:

La primera réplica reposa sobre el siguiente argumento de la réplica de los *sistemas*:

> Aun cuando es cierto que el individuo encerrado en la habitación no entiende el relato, el hecho es que él es sólo una parte del sistema completo que sí entiende el relato. El individuo tiene ante sí un enorme libro en el que están escritas las reglas, cuentan con una gran cantidad de papel y lápices para realizar los cálculos, tiene “bancos de datos” de los conjuntos de símbolos chinos. Entonces, la comprensión no puede atribuirse sólo al individuo, sino a todo el sistema del que él es solo una parte. (Searle, 1994, p. 88)

No obstante, para Searle esta réplica no ataca su argumento acerca de la comprensión, y afirma que, aunque el sujeto memorizara todos los símbolos chinos e hiciera los cálculos correspondientes y los integrara todos al sistema, seguiría careciendo de comprensión, y esta termina siendo necesaria para el sistema, de tal modo que el sistema terminaría siendo solo una parte. De tal forma que “todo lo que se sabe es que los distintos símbolos formales ingresan por un extremo, se manipulan conforme a reglas escritas en inglés y otros símbolos salen por el otro extremo” (Searle, 1994, p. 89), en resumidas palabras, “el propósito del ejemplo original era argumentar que esa manipulación de símbolos por sí misma no basta para comprender chino en ningún sentido literal, porque el hombre podría escribir ‘guara guara’ después de ‘güiri güiri’ sin entender nada del chino” (Searle, 1994, p. 89).

Una segunda réplica que se le hace a Searle es la del “robot”:

> Supongamos que ponemos una computadora dentro de un robot y que ésta no solo acepta símbolos formales como entrada y produce símbolos formales como salida, sino que efectivamente opera al robot de tal manera que éste haga algo muy parecido a percibir, caminar, desplazarse, martillar, comer, beber o lo que quiera. El robot tendría, por ejemplo, una cámara de televisión integrada la cual le permitiría “ver”; tendría brazos y piernas para

> poder “actuar”, y todo esto lo controlaría su cerebro de computadora. A diferencia de la computadora de Schank, este robot sería capaz de una genuina comprensión y de otros estados mentales. (Searle, 1994, p. 91)

Antes de mirar la respuesta que da Searle a dicha réplica, se mirará brevemente en qué consiste el programa de Schank:

Este programa representa la información que se tiene acerca del mundo; su función consiste en dar respuesta a preguntas de su entorno teniendo en cuenta una serie de libretos que han sido creados para responder de forma adecuada en diversas situaciones. Dicho programa pretende simular la capacidad humana para comprender relatos y explicar la capacidad humana de comprender relatos y responder preguntas.

> Así pues, por ejemplo, suponga que escucha la siguiente historia: “Un hombre entró a un restaurante y ordenó una hamburguesa. Cuando se la sirvieron, estaba totalmente quemada, así que el hombre estalló en cólera y abandonó el restaurante furioso, sin pagar la hamburguesa ni dejar propina” Ahora bien, si le preguntamos: “¿se comió el hombre la hamburguesa?”, usted probablemente respondería: “No, no se la comió”. De igual manera, si usted escucha lo siguiente: “Un hombre entró a un restaurante y ordenó una hamburguesa. Cuando se la sirvieron le agradó mucho. Al salir del restaurante, le dejó a la mesera una generosa propina antes de pagar la cuenta. “si se le pregunta: ¿se comió el hombre la hamburguesa?”, usted probablemente responderá: “Sí, se la comió. (Searle, 1994, p. 83)

Las máquinas de Schank pueden responder preguntas similares acerca de restaurantes formuladas en estos términos. El programa cuenta con una “representación” del tipo de información que poseen los seres humanos acerca de los restaurantes, de forma tal que las respuestas que arroja el programa son correctas. Sin embargo, dicho programa carece de comprensión, porque está diseñado para seguir un libreto en el que solo debe arrojar preguntas del tipo que un ser humano daría en un restaurante.

Según Searle esta réplica permite aceptar que la cognición no solo se limita a una manipulación de símbolos como usualmente se cree, sino que es evidente, que dicha réplica involucra relaciones causales del mundo exterior. No obstante, el hecho del que robot pueda desplazarse, caminar o parecer percibir como se menciona en el experimento anterior, no quiere

decir que este posea entendimiento o intencionalidad. En palabras de Searle (1994) “al ejemplificar concretamente el programa yo no tengo estados intencionales pertinentes: todo lo que hago es seguir instrucciones para manipular símbolos formales” (p. 92).

Una tercera réplica es la del *simulador de cerebros,* donde se diseña un programa que no representa la información al estilo de un libreto, sino que se simula la secuencia real de emisiones neuronales:

> La máquina no opera solo con un programa serial, sino con todo un conjunto de programas que operan en paralelo, de la misma manera en que supuestamente operan los cerebros humanos reales cuando procesan el lenguaje natural. Ahora bien, no cabe duda de que en este caso tendríamos que afirmar que la máquina ha comprendido los relatos. (Searle, 1994, p. 92)

Aunque existe una entrada del chino y la simulación de la estructura de la sinapsis del cerebro chino, y del mismo modo, existe una salida o una respuesta del chino, no existe una comprensión como tal del chino. Según Searle (1994), “el problema con el simulador del cerebro es que simula aspectos equivocados de este” (p. 93). A lo que quiere llegar con dicha réplica, es que si se simula la estructura formal del cerebro, se sigue presentando el mismo problema, no hay una simulación de estados Intencionales, por ejemplo, lo que imposibilita el entendimiento y la comprensión. De ahí que sea necesario un estudio riguroso sobre los procesos y capacidades causales que residen en nuestra estructura neurofisiológica.

Una cuarta réplica es la de la *combinación,* formulada en Berkeley y Stanford:

> Imagínenos un robot que cuenta con una computadora en forma de cerebro alojada en su cavidad craneal; imaginemos que la computadora está programada con todas la sinapsis de un cerebro humano; imaginemos también que todo el comportamiento del robot es indistinguible del de un ser humano, e imaginemos ahora todo esto como un sistema unificado y no solo como una computadora con entradas y salidas de datos. Sin duda, en tal caso tendríamos que atribuirle intencionalidad al sistema. (Searle, 1994, p. 94)

Según Searle, el robot terminaría manipulando símbolos y las salidas de información seguirían siendo perfectas. En efecto, no habría distinción alguna entre un ser humano y un computador, no obstante, el problema radicaría en la intencionalidad, ya que por más que simule la conducta humana, esto se reduce a una simulación puramente formal, y no se le podrían atribuir

estados Intencionales. Tener una mente, según Searle, va más allá de tener procesos formales o sintácticos; además, los estados mentales tienen cierto tipo de contenidos. “Si mis pensamientos se me presentan en cadenas de símbolos tiene que haber más que cadenas abstractas, puesto que las cadenas por sí mismas no pueden tener significado alguno” (Searle, 1995, p. 37). De forma tal que, si los pensamientos son sobre algo, entonces han de tener un significado, y los programas funcionan de forma puramente sintáctica, en este punto radica la diferencia entre un programa de computador que intenta simular la mente humana.

Una quinta y última réplica, es *la réplica de las otras mentes*, formulada desde Yale: “Como la computadora puede aprobar (en principio) las pruebas de conducta tan bien como cualquiera; por tanto, si se ha de atribuir cognición a otras personas, también debe atribuírsele a las computadoras” (Searle, 1994, p. 96). Una de las respuestas que ofrece Searle se sustenta en que los resultados de cómputo pueden existir sin cognición, adicional a eso, cuando se alude a las ciencias cognitivas estas presuponen la existencia de pensamiento, confiadas en que este es una manipulación de símbolos con contenido.

Una de las conclusiones generales que se desprende de las réplicas mencionadas, es que los programas se definen en términos puramente formales o sintácticos y, dado que la mente tiene un contenido mental intrínseco, se sigue que el programa no puede por sí mismo constituir una mente. “El argumento descansa en la simple verdad lógica de que la sintaxis no es lo mismo que, ni es por sí misma suficiente para, la semántica” (Searle, 1996, p. 205).

Así mismo, las conclusiones que ha ofrecido Searle acerca del argumento de la habitación china, le han permitido hacer las siguientes aseveraciones. En un primer momento, llegó a la conclusión de que la IA en sentido fuerte hace afirmaciones falsas. Una de ellas reposa sobre la idea de que la mente es algo más que una parte del mundo biológico natural y que esta se puede describir de manera puramente formal, negando de esta manera la naturaleza biológica de la mente. “Los procesos mentales que nosotros consideramos que constituyen una mente son causados, enteramente causados, por procesos que tienen lugar dentro del cerebro” (Searle, 1995, p. 45).

Otra de las afirmaciones falsas sobre la IA tiene que ver con hecho de concebir la mente como algo puramente sintáctico; mientras que la conciencia, los pensamientos, sentimientos y emociones constituyen los procesos y estados mentales; la computadora, por el contrario, se define únicamente en términos de su capacidad para llevar a cabo programas, los cuales se especifican de

manera puramente formal, sin poseer contenidos semánticos. “Por definición el computador es incapaz de *duplicar* esos rasgos por muy poderosa que pueda ser su capacidad para *simular*” (Searle, 1995, p. 43).

Una última afirmación que hace Searle sobre la IA en su libro *Mentes, cerebros y ciencia*, destruye por completo la idea de intentar crear una mente. Según el filósofo, “el proyecto de intentar crear mentes diseñando solamente programas está condenado a muerte desde el principio” (Searle, 1995, p. 46), ya que el resultado termina siendo algo puramente formal, además, “las propiedades computacionales del cerebro simplemente no bastan para explicar su funcionamiento para producir estados mentales” (Searle, 1995, p. 46).

Ahora bien, si se miran las obras posteriores del filósofo se constata que los argumentos y las conclusiones que expone la IA en sentido fuerte, terminan siendo un sinsentido; si en un primer momento afirmó que eran un error, posteriormente las consideró de tal manera. Pero antes de mirar la razón por la cual Searle consideró a la IA en sentido fuerte como un sinsentido, se verán algunas consideraciones que expone Searle sobre la ciencia cognitiva y la computación, en su libro *El redescubrimiento de la mente*.

Uno de los mayores inconvenientes de la ciencia cognitiva es que esta se fundamenta sobre supuestos que en su mayoría son erróneos. En realidad, el estudio del cerebro y la conciencia no tienen mayor interés para la ciencia cognitiva y, aun así, los mecanismos cognitivos están implementados en el cerebro. Además, “los procesos que explican la cognición son inconscientes no sólo de hecho, sino en principio” (Searle, 1996, p. 202). En este sentido, los procesos mentales cognitivos, en su mayoría, son inconscientes y son computacionales. Afirmación que Searle critica notablemente, y para ello alude a un ejemplo concreto: SOAR[14] es un tipo de arquitectura de ordenador, y uno de los programas que está incorporado en SOAR es un robot que cumple determinadas funciones y responde a las órdenes que se le dan:

> Así, por ejemplo, el robot responderá adecuadamente a la orden: “Selecciona un bloque con forma de cubo y muévelo tres espacios a la izquierda”. Para hacer esto, el robot tiene tanto unos sensores ópticos como unos brazos, y el sistema funciona porque implementa un conjunto de manipulaciones formales de símbolos que están conectadas con

[14] SOAR es un sistema desarrollado por Alan Newell y sus colegas en la Universidad de Carnegie Mellon. El nombre es un acrónimo de Stale, Operator, Ami Result. Para una exposición, véase Waldrop (1988).

> transductores que reciben inputs de los sensores ópticos y envían outputs a los mecanismos motores. (Searle, 1996, p. 203)

Para que un ser humano pueda desarrollar todas las funciones que se mencionan en el ejemplo anterior, debe ser consiente, de lo contrario no podría ni siquiera mover los bloques, ya que para ello se necesita dicho rasgo y, desde luego, la intencionalidad.

Ahora bien, ¿por qué el cognitivismo ha sido tan atractivo para diversos académicos y pensadores, si hasta el momento se han vislumbrado varios errores que lo dejan sin fundamentos sólidos?

> Comenzamos con dos resultados de lógica matemática, la tesis de Church-Turing y el teorema de Turing. Para nuestros propósitos, la tesis de Church-Turing enuncia que para cualquier algoritmo hay alguna máquina de Turing que puede implementar el algoritmo. La tesis de Turing dice que hay una máquina universal de Turing que puede simular cualquier máquina de Turing. Ahora bien, si los ponemos juntos, obtenemos el resultado de que una máquina universal de Turing puede implementar un algoritmo cualquiera. (Searle, 1996, p. 207)

El cognitivismo se mostró atractivo porque, en cierto modo, algunas capacidades mentales humanas son algorítmicas, es decir, la resolución de una operación básica o compleja recorre los pasos de un algoritmo. De modo tal que, tanto el ordenador mecánico como el ser humano pueden implementar el mismo algoritmo para resolver determinada operación matemática. Es necesario aclarar que, si bien el ordenador humano recorre los pasos de un algoritmo, lo hace de modo causal y lógico: "lógico porque el algoritmo proporciona un conjunto de reglas para derivar los símbolos de *output* a partir de los símbolos de *input*, y causal porque el agente está haciendo un esfuerzo consciente para recorrer todos los pasos" (Searle, 1996, p. 224), en el caso del ordenador mecánico, el funcionamiento del sistema incluye un homúnculo externo.

Una conclusión que se desprende de este argumento, radica en la simulación, ya que, si es posible simular el proceso de realizar una operación matemática, también se podrían simular procesos tales como comprender un lenguaje, tener una percepción visual, etc.

El problema grueso que atañe esta discusión radica en la semántica, ya que los programas se definen en términos puramente sintácticos. Y la teoría de la demostración dio a conocer que las

relaciones semánticas entre proposiciones pueden reflejarse enteramente por medio de las relaciones sintácticas.

> Supongamos ahora que los contenidos mentales que están en la cabeza se expresan sintácticamente en la cabeza; entonces todo lo que necesitaríamos para dar cuenta de los procesos mentales serían procesos computacionales entre los elementos sintácticos que están en la cabeza. (Searle, 1996, p. 209)

No obstante, este tipo de razonamiento termina siendo algo problemático, y una de las razones consiste en las confusiones y la poca claridad que hay respecto al funcionamiento del cerebro e incluso cuestiones fundamentales como ¿Qué es un ordenador digital? ¿Qué es exactamente un algoritmo? Etc.

Según Searle, si se toma en serio la idea de que el cerebro es un ordenador digital "lo que obtenemos es el poco confortable resultado de que podríamos hacer un sistema prácticamente de cualquier cosa que haga precisamente lo que el cerebro hace" (Searle, 1996, p. 212), lo que genera un sin número de dificultades.

Una primera dificultad a la que se enfrenta el cognitivismo es que la sintaxis no es intrínseca a la física. Ahora bien, para los defensores del computacionalismo los ordenadores pueden hacerse de un rasgo indefinido de *hardware*, además, se definen sintácticamente en términos de asignaciones de ceros y unos, no obstante, de este razonamiento se derivan algunas consecuencias que para Searle terminan siendo desastrosas:

> 1. Si la computación se define en términos de la asignación de sintaxis, entonces todo puede ser un ordenador digital, puesto que a cualquier objeto se le podrían hacer adscripciones sintácticas. Se podría describir cualquier cosa en términos de ceros y unos.
>
> 2. Peor aún, la sintaxis no es intrínseca a la física. La adscripción de propiedades sintácticas es siempre relativa a un agente u observador que trata como sintácticos ciertos fenómenos físicos. (Searle, 1996, p. 213)

El hecho de que la sintaxis no sea un rasgo intrínseco físico, sino que esté sujeta a una noción relativa del observador, impide que se pueda tomar en serio el problema de la realizabilidad universal, dado que los estados computacionales no se descubren dentro de la física, sino que se le

asignan a la física. “La realizabilidad múltiple de los procesos computacionalmente equivalentes en diferentes medios físicos no es sólo una señal de que los procesos son abstractos, sino de que no son en absoluto intrínsecos al sistema. Dependen de una interpretación desde fuera” (Searle, 1996, p. 214).

Por ello, a continuación, se volverá a la discusión inicial sobre las conclusiones a las que Searle llegó en un segundo momento sobre la IA en sentido fuerte. El filósofo no se había percatado del sinsentido en el que cae la IA en sentido fuerte; si se aplica al modelo computacional la caracterización de un proceso como computacional, dicho proceso computacional no identifica un rasgo intrínseco de la física, sino que termina siendo una caracterización relativa al observador. En palabras de Searle: “No hay ninguna manera en que pueda descubrirse que algo es intrínsecamente un ordenador digital, puesto que su caracterización como ordenador digital es siempre relativa a un observador que asigna una interpretación sintáctica a los rasgos puramente físicos del sistema” (Searle, 1996, p. 215).

Vale la pena hacer una distinción entre lo que son rasgos intrínsecos y rasgos relativos al observador. Expresiones como masa, extensión, gravedad y volumen, nombran rasgos del mundo que son intrínsecos, independientemente de la existencia del observador; a diferencia de los rasgos relativos al observador, los cuales existen gracias a la interpretación del mismo. La sintaxis no es intrínseca a la física, sino que depende del observador, en ese sentido, es válido afirmar que rasgos computacionales son intrínsecos al observador.

Otra dificultad con la que se tropieza Searle tiene que ver con la *falacia del homúnculo*. “La sintaxis no es parte de la física” (Searle, 1996, p. 217). De tal manera que, si la computación se define en términos puramente sintácticos, entonces nada es un ordenador digital, ya que esto termina siendo relativo al observador como se ha expuesto en los argumentos anteriores.

Otra dificultad que se suma tiene que ver con que la sintaxis no tiene poderes causales. Es común, sobre todo en ciencias naturales ver explicaciones causales de diferentes fenómenos. Según Searle, el cognitivismo aparentemente ofrece una explicación causal: “Se supone que los mecanismos por los que el cerebro produce cognición son computacionales, y al especificar los programas hemos especificado las causas de la cognición” (Searle, 1996, p. 220). Además, el cognitivismo se sustenta bajo la tesis de que existe una gran cantidad de símbolos, ceros y unos, que causan la cognición. No obstante, esta afirmación carece de sentido, ya que los ceros y lo unos

tienen existencia relativa al observador, en palabras de Searle: “El programa implementado no tiene poderes causales distintos del medio que lo implementa puesto que el programa no tiene existencia real, no tiene ontología más allá del medio que lo implementa” (Searle, 1996, p. 220).

Lo que quiere decir es que no existe una causación intrínseca al sistema como sucedería como un “ordenador humano”, en el cual se da un cumplimiento de reglas de manera consciente, por el contrario, con el ordenador mecánico sucede que no sigue reglas y que una asignación de interpretación computacional es relativa a quien la asigna.

> Ahora bien, el cognitivismo nos dice que el cerebro funciona como el ordenador comercial y que esto causa la cognición. Pero sin un homúnculo, tanto el ordenador comercial como el cerebro tienen únicamente modelos y los modelos no tienen poderes causales adicionales a los que tienen los medios que los implementan. De esta manera, parece que no hay ningún modo en el que el cognitivismo pueda dar una explicación causal de la cognición. (Searle, 1996, p. 221)

Sin embargo, una explicación causal ordinaria sostiene que cuando una máquina está implementando un programa, este arroja unas respuestas de manera causal. Los defensores del cognitivismo aluden a la simulación, ya que la idea es programar un ordenador que simule alguna capacidad cognitiva, y en ese sentido, se podría hablar de una reacción que se genera de manera causal. Pero tampoco es correcto afirmar que porque existe una simulación entonces se dan procesos causales como tal. “El único sentido en el que la especificación del modelo proporciona por sí mismo una explicación causal es este: si se sabe que existe cierto patrón en un sistema, se sabe que hay cierta causa que es responsable del patrón” (Searle, 1996, pp. 223-224).

CAPÍTULO IV

Críticas al argumento de la habitación china de John Searle

El argumento y experimento mental de la habitación china de John Searle ha sido tan famoso en los argumentos de la filosofía contemporánea, que se ha mantenido siempre en un rango amplio de discusión. Diferentes académicos y estudiosos de teorías cognitivas y computacionales han sostenido una postura crítica frente a este argumento, el cual ha intentado socavar los fundamentos de las mismas. Para empezar, se vislumbrarán algunas de las réplicas que se le han hecho al argumento de la habitación china y que, ya en el capítulo anterior han sido mencionadas, así como las repuestas que Searle ha ofrecido a estas; todo esto, con la intención de dejar ver los problemas generales que ha suscitado el argumento como tal.

Aunque en el capítulo anterior se describe con detalle el argumento de la habitación china tal como Searle lo publicó por primera vez en *Mentes, cerebros y programas*, es necesario redondear el argumento para tener presente la conclusión a la que el filósofo llegó y sobre la cual ha rondado un sinnúmero de críticas.

> Searle se imagina a sí mismo solo en una habitación siguiendo un programa de computación para responder a caracteres chinos que le son tirados bajo la puerta. Searle no sabe nada de chino y, sin embargo, por el seguimiento del programa, por la manipulación de símbolos y números que un computador hace, él manda sus propias líneas de caracteres chinos bajo la puerta y, estos llegan a los que están afuera, lo que los hace suponer que hay un hablante del chino en la habitación. (N. del T.) (Cole, 2020)

La conclusión estrecha del argumento sostiene que, aunque el computador digital pareciera como si comprendiera un lenguaje al arrojar las respuestas en chino, en efecto, no sucede de tal forma, ya que los computadores solo usan reglas sintácticas para manipular símbolos, pero no existe una comprensión del significado de las palabras como tal. La conclusión más amplia del argumento, sostiene que “las mentes deben resultar de procesos biológicos; los computadores pueden en el mejor de los casos simular esos procesos biológicos” (N. del T.) (Cole, 2020), de esta conclusión se despliegan una serie de implicaciones semánticas, filosóficas, computacionales, etc. que harán parte del grueso de la discusión en el presente capítulo.

La IA ha demostrado “capacidades” que simulan o, incluso, parecen superar la inteligencia humana a la hora de manipular juegos, o sostener conversaciones en un lenguaje natural en función del servicio al cliente, por ejemplo. Por su parte Alan Turing, uno de los pioneros en IA, estaba convencido de que los computadores excederían la inteligencia humana, además, los primeros defensores de la IA argumentaban a favor de la comprensión de algunos lenguajes naturales por parte de una computadora digital. Todas esas aseveraciones han sido fuertemente criticadas por Searle, quien considera absurdo que un computador digital pueda comprender el lenguaje o pensar.

Décadas después, cuando Searle describe la conclusión de su argumento en términos de conciencia e intencionalidad, afirma:

> La implementación de un programa de computador, no es por sí suficiente para la conciencia o la intencionalidad. La computación está definida en términos puramente formales o sintácticos, mientras que las mentes tienen reales contenidos mentales semánticos y no podemos llegar de lo sintáctico a lo semántico solo teniendo las operaciones sintácticas nada más (N. del T.) (Cole, 2020)

Con este argumento, Searle deja claro que correr un programa o arrojar respuestas específicas no afecta la comprensión en absoluto. “Un sistema, yo, por ejemplo, no podría adquirir una comprensión del chino solo por atravesar los pasos de un programa de computador que simula conducta de un hablante de chino” (N. del T.) (Cole, 2020).

4.1. Réplicas al argumento de la habitación china

4.1.1. Réplica de los sistemas.

En esta réplica se argumenta que el hombre que se encuentra en la habitación, es solo una parte de todo el sistema; analógicamente, como la CPU, o como la unidad de procesamiento central que tan solo hacen parte de un sistema más grande; ya que este está formado por otros componentes que incluyen una base de datos, una memoria, una serie de instrucciones, etc. De forma tal que es el sistema en su totalidad el que responde a las preguntas en chino. Ahora bien, si se vuelve al hombre que se encuentra en la habitación, este sería únicamente una parte del sistema ya que a su alrededor hay otros componentes como un libro de reglas, papel, lápices para hacer cálculos, y un conjunto de símbolos chinos. En resumidas palabras: “ La comprensión no puede atribuirse sólo al individuo, sino a todo el sistema del que él es solo una parte” (Searle, 1994, p. 88).

Uno de los primeros en publicar la réplica de los sistemas fue Ned Block, acompañado de otros académicos, incluido Georges Rey, quien para el año 1986 afirmó que el hombre de la habitación del experimento de Searle es solo una CPU del sistema, así mismo Ray Kurzweil en el año 2002, sostuvo que el ser humano es solo un implementador, y que las propiedades del implementador no son necesariamente las propiedades del sistema, además, se considera un defensor del test de Turing convencido de que si "un sistema despliega la aparente capacidad de comprender chino, él debería de hecho entender chino" (N. del T.) (Cole, 2020). En otras palabras, afirman que Searle cae en una contradicción al afirmar que la máquina puede hablar y arrojar respuestas en chino, pero no comprender el chino.

Searle responde a esta réplica de una manera sencilla. Si se imagina que el individuo memoriza todos los símbolos chinos, todas las instrucciones y la base de datos que hay en la habitación (traslada todo el sistema a su mente); fuera de ella pude hablar chino y responder a lo que se le pregunta, sin embargo, seguiría careciendo de comprensión, ya que no tendría forma de adherirle algún significado a los símbolos formales que ha memorizado, no obtendría semántica.

Ahora ser verá un ejemplo del sistema extendido:

> Si Otto, que sufre pérdida de memoria puede recuperar sus habilidades de memoria externalizando algunas de las informaciones a su cuaderno de notas, entonces, Searle podría hacer lo contrario, internalizar las instrucciones y los cuadernos de notas, de esta manera, él adquiriría las habilidades que tenía el sistema extendido. (N. del T.) (Cole, 2020)

Si bien Searle concluye afirmando que así haya una internalización de las instrucciones y los cuadernos de notas, no hay comprensión de la palabra, por ejemplo, no se sabe lo que significa la palabra china para hamburguesa, no se da la semántica.

Frente a la réplica de los sistemas y a la respuesta que ofrece Searle, se ha argumentado que si "el operador de la habitación memoriza las reglas y hace todas las operaciones dentro de su cabeza, el operador de la habitación no se convierte en el sistema", tal como lo pretendía Searle cuando responde a esta réplica. Así, el individuo de la habitación, no podría ser el sistema, sino una subparte de él, como se ha venido argumentado en líneas anteriores.

> En el caso de la habitación china, una persona es un monóglota inglés y el otro es un monóglota chino. El total desconocimiento del significado de la persona que habla inglés

> de las respuestas chinas, no muestra que ellas no son comprendidas, esta línea de distintas personas nos lleva a la réplica de la mente virtual. (N. del T.) (Cole, 2020)

4.1.2. Réplica de la mente virtual.

El argumento de la mente virtual sostiene que el operador de la habitación china no comprende chino únicamente por correr o instanciar la máquina de papel, en dicha réplica se argumenta que lo importante es que la comprensión es creada, "ya que sostiene que un sistema que corre un programa puede crear nuevas entidades virtuales que son distintas del sistema como un todo y también de los subsistemas, tales como la CPU, o el operador" (N. del T.) (Cole, 2020).

Quienes argumentan a favor de la réplica de la mente virtual, consideran que el error del argumento de la habitación china radica en afirmar que, desde un punto de vista de la IA fuerte, el computador o el sistema entiende chino; cuando la afirmación acorde, desde un punto de vista de la IA, debería ser que el computador que corre un programa crea comprensión del chino.

Algunos estudiosos de la IA han afirmado que, por ejemplo, un modelo familiar de agentes virtuales son caracteres de juegos o videos de computación, los cuales tienen habilidades y personalidades; estos caracteres no son idénticos con el *hardware* o el sistema que los crea. Así

> Un solo sistema instanciador puede controlar dos agentes distintos o físicos simultáneamente, uno de los cuales conversa solo en chino y, él otro conversa solo en inglés, se manifiestan personajes diferentes, memorias y habilidades cognitivas, por lo tanto, esta réplica nos pide distinguir entre las mentes y sus sistemas realizadores. (N. del T.) (Cole, 2020)

> Tim Maudlin, considera a los sistemas físicos como aquellos capaces de implementar un sistema computacional corriendo un programa, […] su discusión se da alrededor de su imaginaria máquina Olympia, un sistema de baldes que trasfieren agua implementado una máquina de Turing. El objetivo principal de Maudlin es la afirmación computacionalista de que una máquina tal podría tener conciencia fenomenal. (N. del T.) (Cole, 2020)

En otras palabras, el argumento de Maudlin recae sobre la idea del funcionalismo, en el que un computador se puede hacer de *hardwares* distintos, sin necesidad de que haya unas propiedades físicas determinadas o específicas, porque en últimas, lo que importa es que los programas corran y que los *inputs* y lo *output* funcionen de manera correcta. En este sentido y

siguiendo la línea de argumentación de Maudlin, no es necesario un ser biológico hecho con un *hardware* biológico para que se genere pensamiento.

David Cole por su parte, argumenta en contra del argumento de la habitación china y afirma que la incapacidad de Searle para comprender chino en la habitación, no demuestra que no haya una comprensión creada. El hecho de que Searle no comprenda el chino mientras opere la habitación, no muestra que la comprensión no esté siendo creada.

En conclusión, la réplica de la mente virtual sostiene que Searle se justifica bajo la premisa de que no se da una comprensión de chino, para no admitir que él podría entender chino en la habitación. De tal manera que el argumento de la habitación china no puede refutar la afirmación de que la IA es incapaz de asumir la posibilidad de crear comprensión usando un computador digital programado.

4.1.3. Réplica del simulador de cerebros.

Usualmente los programas de IA operan bajo guiones y libretos sobre cadenas de oraciones o símbolos, sin embargo, la réplica del simulador de cerebros supone la simulación de la secuencia real de emisiones neuronales que ocurren en el cerebro de un hablante nativo del chino. El computador no solo opera bajo un programa serial, sino que reúne todo un conjunto de programas que procesan el lenguaje natural, tal como lo hacen los cerebros humanos. “En la medida en que el computador trabaja de la misma forma que el cerebro de un hablante nativo del chino, el procesamiento de información se dará de la misma manera y entenderá el chino” (N. del T.) (Cole, 2020).

Serle responde a esta réplica con un contraejemplo:

> Supóngase que, en la habitación, el hombre tiene un gran conjunto de válvulas y tuberías, arregladas de la misma forma que las neuronas en un cerebro de un hablante nativo del chino, el programa ahora le dice al hombre cuáles válvulas abrir en respuesta a que *input*, Searle afirma que es obvio que no habría comprensión del chino. (N. del T.) (Cole, 2020)

Además, concluye que una simulación no podría ser la cosa real. En su libro *Mentes, cerebros y ciencia*, considera absurdo que se crea que una simulación computacional de procesos mentales tenga efectivamente procesos mentales, y alude a varios ejemplos tales como que un simulador de un incendio, no es el incendio real o, un simulador de una tormenta no es la tormenta

real, la pregunta que inquieta a Searle es ¿Por qué se habría de creer entonces que una simulación de procesos mentales tiene realmente procesos mentales?

En el año 1980, Pylyshyn se inquieta por los sistemas híbridos, y escribe que si las células del cerebro fueran reemplazadas por circuitos integrados de chips programados de una forma tal que se mantuvieran los *input-output* idénticos a la reemplazada, se podrían lograr procesos mentales reales. Así mismo, Rey en 1986, argumenta que resulta razonable atribuirle intencionalidad a un sistema tomándolo como un todo.

> Searle está de acuerdo en que, de hecho, podría ser razonable atribuir comprensión a un sistema androide, pero únicamente en la medida en que usted no sepa cómo trabaja, tan pronto como usted conozca la verdad que es un computador que manipula símbolos sin comprensión sobre la base de la sintaxis y no el significado, usted dejará de atribuirle intencionalidad a este. (N. del T.) (Cole, 2020)

Frente al argumento que expone Searle, los Churchlands (Paul y Patricia Churchland) coinciden con el filósofo, en cuanto a que el hombre de la habitación china realmente no comprende chino, pero sostienen que el argumento en sí mismo, además, de explotar la ignorancia de los fenómenos cognitivos y semánticos, deja por fuera la teoría conexionista[15] (véase capítulo I). Ellos por su parte, argumentan que el cerebro es como un sistema conexionista y no como un sistema que manipula símbolos, por lo que sostienen que el sistema de la habitación china usa estrategias computacionales erróneas, lo que debilita enormemente el argumento de la habitación china.

Otro crítico de Searle, Andy Clark, argumenta en su libro *Microcognition,* que está de acuerdo con Searle cuando afirma que un computador con solo correr un programa de Schank no alcanzará comprensión como tal, no obstante, el filósofo se equivoca al pensar que ningún fenómeno computacional puede generar comprensión. Sus críticos, aseveran que Searle está equivocado acerca de los modelos conexionistas, y que incluso los deja por fuera. Por su parte,

[15] El modelo neuronal o conexionista aparece alrededor de la década de los años 80. Este modelo pretende alcanzar una mayor cercanía con la estructura del cerebro, a diferencia del modelo simbólico que centra su atención en la sintaxis y el leguaje como punto de partida. “En los modelos conexionistas los contenidos mentales no se codifican ya en fórmulas sintácticas, sino en redes de actividad. Por tanto, si nuestra mente fuese una red conexionista, entonces parece que el modo como se instancian los contenidos mentales no podría ser, como propone la Imagen Sintáctica, cerebro - sintaxis => semántica, sino redes cerebrales - realizan redes conexionistas redes conexionistas = codifican => contenidos mentales” (Corbí y Prades, 2007, p. 162).

Clark sostiene que el cerebro piensa en virtud de sus propiedades físicas, y las propiedades más importantes para Clark, son las estructuras variables y flexibles, que, aunque no las tienen los sistemas convencionales de IA, no se sigue que, por ello, el computacionalismo y el funcionalismo sean falsos. Más bien, "se debería buscar en una descripción funcional de grano fino, a un nivel de conjuntos neuronales" (N. del T.) (Cole, 2020).

4.1.4. Réplica de las otras mentes.

Esta réplica está relacionada con la anterior:

> ¿Cómo sabe usted que las otras personas comprenden chino o algo más? Solo por su conducta. Ahora el computador puede pasar el test conductual también como ellos pueden en principio. Así, si usted va a atribuirle cognición a otra persona, usted debería en principio atribuírsela a los computadores. (N. del T.) (Cole, 2020)

De la misma forma que en las ciencias cognitivas se presupone la realidad y el conocimiento de lo mental, en las ciencias físicas se tiene que presuponer la realidad y cognoscibilidad de los objetos físicos.

Pese a que las presuposiciones que se pueden hacer sobre la conducta de los seres humanos no son tan relevantes, y la razón es que muchas veces terminan siendo falsas. Ahora bien, si estas presuposiciones que se hacen sobre la conducta de los seres humanos son pragmáticas, dichas presuposiciones también son válidas para los computadores.

Frente a este argumento, Searle reacciona y expone que la comprensión es algo más que solo disposiciones complejas conductuales, además, en las conclusiones a las que llega años después de haber publicado el argumento de la habitación china, dejan claro que la intencionalidad requiere de la presencia de estados internos con carácter intrínseco fenoménico, y desde luego, las computadoras no pueden tener intencionalidad intrínseca, sin embargo, esto es rebatible.

Por ejemplo, uno de los intereses de la filosofía reciente, ha centrado su atención en los zombis, criaturas que se ven y se comportan como humanos normales, incluyendo una conducta lingüista, pero no una conciencia subjetiva. No obstante, y pese a que no tienen conciencia subjetiva, no por ello, se puede afirmar que no puede adquirir comprensión. Searle es reacio frente a esa afirmación, y sostiene que, aunque la conducta lingüística es una condición suficiente para atribuir condición a los seres humanos, debe compartir nuestra biología. Sobre esto:

> Hans Moravec, director del laboratorio robótico de la universidad de Carnegie Mellon y, autor del libro, *Una máquina para una mente trascendente*, argumenta que la posición de Searle meramente refleja intuiciones de la filosofía tradicional de la mente que están atrasadas con respecto a la nueva ciencia cognitiva. (N. del T.) (Cole, 2020)

Así mismo, otros críticos sostienen que, así como se atribuye pensamiento a las personas con base en su conducta, el test de Turing, en efecto, debe funcionar; ya que este intenta mostrar como un programa de computador le puede hacer pensar a alguien que es otra persona a la hora de sostener una conversación. En consecuencia, no debería haber ningún tipo de reparo a la hora de atribuirle pensamiento a un computador que esté corriendo un programa y que sea capaz de hacerle pensar a alguien que es inteligente.

4.1.5. Réplica de la intuición.

Muchos de los que se han interesado por el argumento de la habitación china, han notado que el argumento parece estar basado en la intuición; la intuición de que el hombre de la habitación o el computador no pueden tener comprensión, sino que cumplen una tarea estrictamente sintáctica de manipulación de símbolos. A pesar de ello, basar un argumento en intuiciones lo puede debilitar fácilmente.

Block en 1980 argumentó sobre esta réplica y aseguró que, en un primer momento, las intuiciones debían ser abandonadas, y que además, era necesario "llevar nuestro concepto de la comprensión en línea con la realidad de que ciertos robots computadores pertenecen a la misma clase natural de los humanos" (N. del T.) (Cole, 2020). Margaret Boden no está muy alejada de la reflexión que hace Block, y en el año 1988, señala que no se puede en las intuiciones acerca de cómo la mente depende de la materia, ya que los desarrollos de la ciencia los pueden cambiar, "de hecho, la eliminación de la confianza en nuestras intuiciones fue precisamente lo que motivó a Turing para proponer el test de Turing, un test que es ciego al carácter físico de los sistemas que responde nuestras preguntas" (N. del T.) (Cole, 2020).

Algunos críticos del experimento mental de Searle sostienen que el filósofo ha partido de la intuición. Por lo que las razones que expone en el argumento de la habitación china sobre la inteligencia, la comprensión y el significado, podrían ser falsas o poco confiables en relación con la ciencia contemporánea. Wakefield sostiene que una explicación computacional no es un análisis de conceptos, "más bien estamos construyendo una teoría científica del significado que puede

requerir revisar nuestras intuiciones. Como una teoría, esta obtiene evidencia de su poder explicativo, no de su acuerdo con intuiciones pre teóricas" (N. del T.) (Cole, 2020). Otros interesados en el tema como Steven Pinker, afirman que Searle únicamente está explorando hechos acerca de la palabra comprender, pero no tiene una definición clara, lo que le impide atribuir comprensión a las máquinas.

En conclusión, la objeción que principalmente se le hace a Searle, es que este fundamenta el concepto de comprensión en intuiciones, y no en un concepto de comprensión más técnico. Además, Pinker sostiene que el experimento del filósofo se basa en intuiciones no analizadas, y esto dificulta que se tome en serio las consideraciones que expone Searle acerca de la imposibilidad de que un computador pueda pensar. "El resultado simplemente puede ser que nuestras intuiciones con relación a la habitación china no son confiables y, por lo tanto, el hombre de la habitación al implementar el programa puede comprender el chino a pesar de las intuiciones contrarias" (N. del T.) (Cole, 2020).

4.2. Otras críticas al argumento de la habitación china

Desde que el argumento de la habitación china de John Searle fue publicado y leído por una amplia comunidad de académicos y estudiosos de teorías filosóficas y cognitivas, numerosos artículos y comentarios han surgido alrededor de dicho argumento.

Para vislumbrar algunas críticas adicionales a las ya reseñadas se verán dos que se han publicado en artículos como el de Patricia Hanna y William Rapaport, para ello, en este punto se utilizará la tesis de pregrado de la Universidad de Caldas de Fernando Alarcón Varela titulada: *John Searle y la tesis de la inteligencia Artificial fuerte: críticas y réplicas*, quien tradujo dichos artículos, y que son materia de investigación en este proyecto.

Patricia Hanna, una filósofa y estudiosa de teorías lingüísticas y filosóficas, publicó un artículo titulado "Poderes causales y cognición" ("*Causal powers and cognition"*), en el que critica la forma en la que Searle ha utilizado algunos conceptos, como algoritmo, programa y, sobre todo la IA. No obstante, el fuerte de su crítica se centra en el argumento de la habitación china.

Según Hanna, el argumento de Searle en contra de la IA fuerte es muy simple; se enfoca y se dirige en contra del trabajo de Roger Schank. Si se recuerda la historia de Schank, que consiste en que un hombre entra a un restaurante y ordena una hamburguesa, cuando se la sirven esta está

quemada, el hombre sale enfurecido. Ahora bien, ¿se comió el hombre la hamburguesa? ¿Pagó la cuenta? La máquina de Schank respondería que no, tal como lo haría un ser humano. En palabras de Searle, esta máquina consiste en simular la conducta humana y responder preguntas. No obstante, no alcanza el carácter comprensivo. Ahora bien,

> Como Searle interpreta el asunto, los defensores de la IA fuerte afirman que en una secuencia tal de pregunta y respuesta la máquina no sólo está simulando una habilidad humana sino también (1) que la máquina puede decirse literalmente que *entiende* la historia y por lo tanto provee respuestas a las preguntas, y (2) que lo que la máquina y su programa *explican* la habilidad humana de entender la historia y contestar preguntas apropiadamente sobre ésta. (Alarcón, 2009, pp. 91-92)

Searle alude a su argumento de la habitación china para responder a los teóricos de la IA. Dicho argumento ya ha sido mencionado en el capítulo anterior. Searle se imagina encerrado en una habitación, se le proporciona un fajo de letras chinas, adicional a eso, se le entrega una serie de instrucciones y reglas que lo capacitan para correlacionar un conjunto de símbolos formales con otro, de tal manera que es capaz de responder a las preguntas que se le hacen fuera de la habitación como si fuese un hablante nativo del chino.

Searle concluye que las respuestas arrojadas al exterior de la habitación, son la causa de la manipulación de símbolos formales, en el que no hay comprensión de las palabras, a pesar de que sus *inputs* y sus *outputs* son indistinguibles de los de un hablante nativo del chino. “Por lo que al chino se refiere, él sólo se comporta como un computador, en tanto que simplemente realiza las operaciones computacionales en elementos formalmente especificados y es de esta manera sólo una instanciación del programa del computador” (Alarcón, 2009, p. 93).

Aunque Searle reconoce que tanto el programa como el computador constituyen el sistema, se sigue careciendo de comprensión (del chino). En consecuencia, el programa que plantea Schank, el cual intenta imitar la conducta humana, no comprende; solo sigue una serie de reglas e instrucciones que le permiten arrojar respuestas indistinguibles a las de un ser humano, tal como sucede con el hombre que se encuentra en la habitación, el cual arroja respuestas en chino indistinguibles a las que daría un hablante nativo del chino. La explicación a estas respuestas, reposa sobre la base de la manipulación formal de símbolos, no sobre el soporte de la comprensión o el significado, por eso ningún computador puede pensar ni comprender.

> Las operaciones computacionales en elementos puramente especificados de manera formal no tienen una conexión interesante con la comprensión o, más generalmente, con la posesión de estados intencionales, en la medida en que estos implican aquel rasgo de ciertos estados mentales mediante el cual están dirigidos a o son sobre objetos o estados de cosas en el mundo. (Alarcón, 2009, p. 94)

Una de las primeras falencias que Hanna nota en la argumentación de Serle, está ligada al uso de conceptos, para esta filósofa norteamericana, Searle posee una maraña de confusiones y errores sobre lo que es la IA, además, no toma conceptos fundamentales como los de *algoritmo* y *programa* de un modo riguroso, incluso, no deja clara la distinción entre ambos conceptos. Por su parte,

> Los investigadores en IA conciben al modelo computacional como un mecanismo mucho más rico de lo que Searle concibe. Hablar de *distinciones cualitativas*, y *representaciones* indica que estos investigadores de IA asumen que el modelo computacional es capaz de modelar procesos interpretativos y/o interpretados, no sólo símbolos no-interpretados (o "formales"). (Alarcón, 2009, p. 95)

Según la filósofa, Searle confunde los programas[16] con algoritmos. "Una característica muy importante de un algoritmo es que este no es la máquina o el lenguaje específico" (Alarcón, 2009, p. 100), empezando que, a diferencia del programa, un algoritmo puede ser caracterizado como cualquier procedimiento sistemático para resolver un problema o implementar un procedimiento. Por ejemplo, "las recetas para pasteles y las instrucciones para montar una bicicleta son algoritmos" (Alarcón, 2009, pp. 99-100). De modo tal que "los programas, que proveen las explicaciones de la cognición humana, no solo no necesitan ser, sino que de hecho no pueden ser, separados de sus realizaciones en las máquinas" (Alarcón, 2009, p. 100).

Otra de las críticas que se le hace a Searle, es que enfoca su argumento en viejos paradigmas e investigaciones de la IA, lo que lo debilita notablemente. Schank por su parte, no aceptaría la limitación y reducción que hace Searle al otorgarle a las máquinas una función únicamente sintáctica. Además, objeta que, si se analiza con detenimiento el experimento mental de Searle,

[16] "En la IA fuerte… lo que importa son los programas, y estos son independientes de su realización en las máquinas… Pero yo no debería estar sorprendido; a no ser que usted acepte alguna forma de dualismo, el proyecto de la IA fuerte no tiene oportunidad… a menos que la mente no sea sólo conceptualmente sino empíricamente independiente del cerebro, usted no puede cumplir el proyecto, ya que el programa es completamente independiente de cualquier realización" MBP (como se citó en Alarcón, 2009, p. 100).

termina siendo incoherente, al afirmar que el comportamiento del hombre que se encuentra en la habitación es indistinguible al de un hablante nativo del chino. Esa afirmación de Searle implica que hay un componente sintáctico y semántico integrado. En palabras de Hanna (como citó en Alarcón, 2009): “Si, por otra parte, restringimos la habitación a las reglas puramente sintácticas, el comportamiento resultante no reflejará la conducta humana” (p. 97).

Según la filósofa, aunque únicamente existan reglas puramente sintácticas y reglas puramente semánticas, estas podrían ser integradas en una unidad de control. Una aproximación chomskyana, por ejemplo, integra sistemáticamente componentes sintácticos, semánticos y fonéticos para hacer posible la interpretación. Igualmente, Schank propone dotar al sistema con un componente semántico, el cual facilitaría la comprensión de las respuestas que puede arrojar una máquina.

Hanna concluye que el experimento de la habitación china, de entrada, cae en un error, y se sirve de las respuestas que ofrece Schank. Cuando el filósofo formula el experimento sobre la base de que en la habitación hay un hombre hablante del inglés que manipula símbolos chinos sin entenderlos, y que este produce una conducta igual a la de un hablante nativo del chino, no es lícito pensar que el hombre no comprende. “Schank argumentaría que Searle se está enfrentado con el siguiente dilema. Si la habitación está para producir un comportamiento indistinguible, nosotros debemos poner por delante un componente sintáctico/semántico integrado” (Alarcón, 2009, p. 97).

Aunque Searle ha defendido su experimento mental bajo la premisa de que el hombre en la habitación se limita a una manipulación formal de símbolos, a algo puramente sintáctico que no compromete lo semántico. La literatura filosófica contemporánea argumenta que, si se reduce el experimento a una explicación puramente formal y sintáctica, el comportamiento del sistema no podrá reflejar una conducta indistinguible de un hablante nativo del chino como lo plantea el mismo Searle en su experimento, por lo que este se queda sin salida.

William Rapaport, especialista en filosofía de la mente y en ciencia computacional, publicó un artículo titulado “El experimento mental de Searle” (“*Searle's experiments with thought”*), Rapaport se considera un crítico del argumento de la habitación china, es más bien, un defensor de la idea de que una máquina, en efecto, puede comprender, así, afirma que existe comprensión en un proceso que es estrictamente sintáctico. Para entender sus afirmaciones acerca de la sintaxis

y la semántica, propuso una serie de experimentos que intentan rebatir el argumento. Uno de ellos, tiene que ver con el de la ecuación algebraica:

Rapaport sabía que había una forma determinada para resolver esta ecuación $2x + 1 = 3$, cuenta que siempre utilizaba el mismo procedimiento para llegar al resultado, de tal manera que los pasos podían ser generalizados y hacerse más precisos; "pero dar cuenta de cada paso es *puramente sintáctico*: la manipulación de símbolos en su más puro sentido" (Alarcón, 2009, p. 111). Asegura Rapaport, que, aunque su procedimiento se reducía a una manipulación formal de símbolos, él entendía cómo resolver tal ecuación.

Sin embargo, se dio cuenta que había otra forma de realizarla, siguiendo unos pasos determinados:

> La técnica presentada era, para mí, radicalmente diferente y muy reveladora. Para resolver la ecuación (1), se decía al espectador que pensara en las ecuaciones como representando una báscula balanceada, con pesos representando las expresiones '$2x + 1$' y '3' en cada uno de los discos de la balanza. (Una balanza moderna fue utilizada para la demostración). Había una restricción: la báscula debía permanecer en balance. (Alarcón, 2009, p. 111)

Aunque él sabía que mediante un procedimiento determinado podría solucionar la ecuación y llegar al resultado, desconocía el significado de las expresiones. Aun así, mediante un proceso sintáctico logró una comprensión semántica. "La pregunta es: ¿Es la comprensión semántica algo cualitativamente diferente de la comprensión sintáctica? […] argumentaré que, aunque tuve una *mejor* comprensión del álgebra, ésta no es una *clase* cualitativamente diferente de comprensión" (Alarcón, 2009, p. 112).

En una primera conclusión, Rapaport asevera que, aunque el hombre de la habitación china desconozca por completo el idioma, y solo cumpla con la función de manipular símbolos, tiene una serie de instrucciones que le permiten adquirir, por lo menos, la comprensión del proceso. En este punto Rapaport siente que el argumento de Searle se debilita, ya que, el hecho de "comprender" el proceso de manipular símbolos dentro de la habitación, implica una comprensión de algo. De forma tal que, aunque el hombre de la habitación no conozca el significado de ninguna de las palabras, sí comprende el proceso para manipular los símbolos y llegar a las respuestas.

Un segundo experimento, propuesto por David Cole, consiste en fusionar el cerebro de Searle con el de Hao Wang, uno es inglés-parlante y el otro comprende el chino. “La finalidad del experimento es suponer que en esta ‘persona fusionada’ hay, en un sentido, ‘partes distintas’: la correspondiente a Wang, capaz de comprender el chino, y la de Searle, capaz de comprender el inglés” (Alarcón, 2009, p. 68). En este caso, Searle terminaría siendo una parte del sistema que no comprende el chino, pero la otra parte del sistema que sería Wang sí lo comprende. Rapaport ofrece una explicación a partir de dos cuestiones: por un lado, a las simulaciones y, por el otro, al componente sintáctico integrado al semántico.

En líneas anteriores se presentaba un argumento en el que Searle consideraba absurdo que una simulación computacional de procesos mentales tuviese efectivamente procesos mentales, y mencionaba ejemplos tales como que una simulación de un incendio no era el incendio real, en consecuencia, la computadora no podría tener procesos mentales tales como comprender y pensar. sin embargo, Rapaport arguye que Searle toma ejemplos inadecuados y, que, por ejemplo, se podría simular un banco de datos, y aunque este como tal no es el banco de datos, no se podría afirmar por ello, que este no cuenta con una información y una base de datos. “A lo mejor *algunas X*’s simuladas, o implementaciones de *X*’s, no son *X*’s (como los ejemplos de Searle (1980, p. 424) de simulaciones de computador de la leche o el azúcar), pero otras lo son” (Alarcón, 2009, p. 113).

> Una forma para argumentar a favor de esto es visualizar un fenómeno mental, así como el pensamiento, como algo abstracto que puede ser implementado en dos medios diferentes; dígase en un cerebro humano y en una computadora. La implementación de la *computadora*, del pensamiento, puede ser tratada como una *simulación* de la implementación humana del pensamiento, y los dos tipos de pensamiento pueden ser distinguidos por diferencias entre los medios que los implementan, a pesar de ser *ambas especies* de pensamiento. (Alarcón, 2009, p. 113)

Respecto al componente semántico, Rapaport plantea otro experimento mental en el que es modificada una parte del cerebro de Searle para que comprenda chino, sin embargo, cuando se le hacen preguntas en inglés a Searle, este las responde en el mismo idioma. “Rapaport se pregunta si la semántica es entendida como un vínculo de los símbolos con el mundo externo, o más bien como una relación de las representaciones internas de, o sobre, los objetos externos” (Alarcón, 2009, p. 70). Searle ha ofrecido respuestas a este planteamiento en réplicas que se le han hecho, y argumenta que, aunque el hombre de la habitación memorice todas las reglas y símbolos chinos

para responder a las preguntas de forma indistinguible a como lo haría un hablante nativo del chino, sigue careciendo de comprensión, por lo que la interacción con el medio tampoco es suficiente para adquirir comprensión.

Una última conclusión de Rapaport al aparente problema sintaxis-semántica que se deprende del argumento planteado por Searle, se fundamenta en que, aunque la semántica no puede ser introducida en un programa de la misma manera que en una mente humana, los "vínculos semánticos no son realmente más que símbolos sintácticos pulsantes. Aún si el programa *de Searle es* insuficiente para la comprensión, un mero programa que pudiera hacer más *podría* comprender" (Alarcón, 2009, p. 116).

Varias de las consideraciones que hacen tanto Hanna como Rapaport tiene grandes repercusiones en el argumento de la habitación china. La primera objeción que detecta Hanna en los argumentos de Schank tiene mucho sentido, al considerar el argumento de Searle incoherente. Si el comportamiento del hombre que se encuentra en la habitación es indistinguible del de un hablante nativo del chino, entonces, debe tener los mismos poderes causales del nativo. ¿Qué hace diferente una conducta que es indistinguible? A los ojos de Searle, el hombre de la habitación manipula símbolos, y no comprende, como lo haría el nativo, en esa línea de argumentación, no se podría hablar de una conducta indistinguible. Adicional a eso, Searle da por sentado que únicamente se le puede atribuir mente a algo que tenga los mismos poderes causales del cerebro, y en ese sentido, se da la comprensión. Pero esto ya ha sido refutado, incluso, en la réplica de las otras mentes, se defiende que para atribuirle mente algo, no tiene que tener los mismos poderes causales del cerebro.

En el siguiente apartado, se tendrá en cuenta varias de las críticas y réplicas que se platean acá, especialmente, la réplica de las otras mentes. Se retomará el concepto de intencionalidad intrínseca, y de la sintaxis y la semántica relativa al observador, para concluir que las creencias y los deseos, también son relativos al observador, y que se puede atribuir mente a todo aquello a lo se le pueda atribuir creencias.

Conclusiones y comentarios finales

Muchos han sido los argumentos con los que Searle ha respondido a las réplicas y objeciones que han atacado su argumento de la *Habitación china*; pero hay uno que merece especial atención, en el que da indicios para que su propio argumento se debilite.

En su libro el *Redescubriendo de la mente*, el filósofo deja claro que la sintaxis no hace parte de un rasgo físico; por el contrario, la sintaxis al igual que la semántica son nociones relativas al observador. Cuando publicó por primera vez el argumento de la *Habitación china*, llegó a la conclusión de que la semántica no era intrínseca a la sintaxis, años después, al retomar el argumento, quiso mostrar que la sintaxis no era intrínseca a la física y que dependía enteramente del observador. En efecto, utilizó este argumento para mostrar que, si un ordenador digital se define en términos puramente sintácticos, ese rasgo termina siendo algo relativo al observador, y no algo intrínseco del ordenador.

Ahora bien, para diferenciar entre los rasgos del mundo que son intrínsecos a la física y los rasgos que son relativos al observador, Searle alude a los siguientes ejemplos, en los que explica y deja claro, cómo en los primeros no se necesita de un observador para hacer posible su existencia, a diferencia de los segundos, sin la cual no existirán.

> Las expresiones “masa”, “atracción gravitatoria” y “molécula” nombran rasgos del mundo que son intrínsecos. Si todos los observadores y usuarios dejasen de existir, el mundo contendría aún masa, atracción gravitatoria y moléculas. Pero expresiones tales como “día precioso para ir a merendar al campo”, “bañera” y “silla” no nombran rasgos intrínsecos de la realidad. Más bien, nombran objetos especificando algún rasgo que les ha sido asignado, algún rasgo que es relativo a observadores y usuarios. Si no hubiese habido jamás usuario u observador alguno, habría con todo montañas, moléculas, masas y atracción gravitatoria. Pero si no hubiese habido nunca ningún usuario u observador, no habría rasgos tales como ser un día precioso para ir a merendar al campo, o ser una silla o una bañera. (Searle, 1996, p. 216)

Teniendo en cuenta la distinción que hace Searle, es posible afirmar que un objeto, por ejemplo, tiene una determinada masa y una determinada composición química. En este caso, el objeto está compuesto de manera y metal; el principal componente de la madera es la celulosa, y

del metal, el hierro. Todos estos rasgos son intrínsecos del objeto, es decir, su existencia no depende de ninguna actitud de un observador. No obstante, cuando se describe como una silla o una mesa, se está determinando un rasgo que, según Searle es relativo al observador. Y la explicación que subyace, tiene que ver con el modo cómo las personas han usado o visto estos objetos, de tal manera que ese rasgo solo existe gracias a un observador. "Los rasgos relativos al observador existen solo en relación con las actitudes de los observadores. A los rasgos intrínsecos les importa un higo los observadores y existen independientemente de ellos" (Searle, 1997, p. 30).

Ahora bien, algunos de esos rasgos que son ontológicamente subjetivos son epistémicamente objetivos, dado que es una cuestión de hecho objetivamente apreciable, y cualquier sujeto inmerso en una cultura sabe la función que estos cumplen. Searle, además considera que, aunque dichos objetos son relativos al observador, los rasgos de los observadores que les permiten crear tales rasgos del mundo, son rasgos intrínsecos. En palabras del filósofo "los rasgos intrínsecos de la realidad son aquellos que existen independientemente de todos los estados mentales, salvo los estados mentales mismos, que son también rasgos intrínsecos de la realidad" (Searle, 1997, pp. 30-31), lo que significa que un rasgo intrínseco del observador tiene que ver con su carácter propiamente mental.

El interés de esta discusión consiste en dejar claro que tanto la sintaxis como la semántica no son rasgos intrínsecos y constitutivos de los sujetos, sino que son rasgos relativos a los observadores. La sintaxis, tiene como función establecer una estructura y una conexión correcta entre sintagmas, y la semántica, tiene la función de otorgar un significado a dichos sintagmas. Una expresión tal como *Hoy es un día perfecto para ir a acampar* es relativa a los observadores, y desde luego, se adquiere en un contexto social y cultural, donde los observadores comparten el mismo concepto, lo que Searle define como rasgos epistémicamente objetivos, y que, desde luego, hacen posible la comprensión y la comunicación.

En líneas anteriores se platearon algunos ejemplos propuestos por Searle para hacer la distinción entre los rasgos del mundo que son intrínsecos y los que son relativos al observador; el filósofo reafirmó que expresiones como las se mencionaron en el párrafo anterior, no podrían existir, sino hubiese un observador o usuario que otorgase una caracterización como tal. Teniendo en cuenta el argumento que ofrece Searle y el propósito inicial de debilitar su argumento de la habitación china, se hace necesario dejar el siguiente planteamiento en remojo: si la sintaxis y la

semántica son rasgos relativos al observador ¿Por qué las creencias, deseos, e intenciones, también no podrían serlo? Se volverá sobre este planteamiento en párrafos posteriores, cuando se dé un vistazo a lo que Searle concibe como intencionalidad intrínseca.

La intencionalidad es para Searle un rasgo de lo mental, y la define como aquella mediante la cual los estados mentales se dirigen a algo o son sobre objetos y estados de cosas del mundo, del mismo modo que los actos de habla representan objetos y estados de cosas del mundo. "Así, como mi enunciado de que está lloviendo es una representación de cierto estado de cosas, mi creencia de que está lloviendo es también una representación del mismo estado de cosas" (Searle, 1992, p. 26), por lo que cada estado Intencional consta de un contenido representativo, o sea, cuando se afirma que una creencia es una representación, se está diciendo que esta tiene un contenido proposicional.

Esa capacidad de representar objetos y estados de cosas del mundo es una extensión de las capacidades biológicas más fundamentales de la mente para relacionar el organismo con el entorno, a través de los estados mentales como las creencias y los deseos. Searle defiende y sustenta una teoría de la intencionalidad intrínseca, ya que "hablar de la intencionalidad de la mente equivale a hablar de una característica que es intrínseca a los estados mentales, que no se deriva de alguna forma previa de intencionalidad" (Moya-Cañas, 2003, p. 34).

Ahora bien, Searle afirma que la especificación del contenido intencional es ya una especificación de las condiciones de satisfacción, las cuales se aplican a los estados Intencionales en los que hay una dirección de ajuste. Así, las creencias y los enunciados pueden ser verdaderos o falsos dependiendo de si se cumplen o no. "*Creo que hoy va a llover",* esta creencia será verdadera, si y solo si, llueve, por lo que la dirección de ajuste en este caso es de mundo a mente; por su parte, los deseos e intenciones no pueden ser ni verdaderos ni falsos, simplemente se cumplen o se satisfacen, por lo que su dirección de ajuste, en este caso, es de mundo a mente. "*Le ordeno que por favor cierre la puerta"*, esta orden puede ser o no ser cumplida.

Según el análisis que hace Moya Cañas (2003) sobre la teoría de la intencionalidad de Searle "queda claro en qué sentido una creencia es intrínsecamente una representación: lo es en cuanto que consiste simplemente en un contenido intencional y un modo psicológico" (p. 35).

> En el ejemplo nombrado, el de la creencia, la dirección de ajuste es de la mente al mundo, pero en el modo psicológico del deseo, ésta es del mundo a la mente. La creencia y el deseo no se pueden separar del contenido representativo, y si el agente es consciente de las condiciones de satisfacción de éstas, es porque están implícitas en las mismas creencias y deseos. Es, insisto, en este sentido en el que el contenido intencional es interno a los estados intencionales. (Moya-Cañas, 2003, p. 35)

En conclusión, la intencionalidad que defiende Searle es intrínseca, ontológicamente subjetiva y no relativa al observador. En párrafos anteriores se planteaba el ejemplo de una silla, y se decía que el uso que el hombre le ha dado es relativo al observador, y aunque Searle está de acuerdo con esa caracterización, considera que la actitud del observador es intrínseca, y lo es precisamente por su estado consciente y, además, porque todos los estados mentales tienen un carácter intrínseco al sujeto. Ahora bien, ¿cómo ese carácter intrínseco que defiende Searle nos permite debilitar su argumento?

Un argumento fuerte, que, al modo de ver de la autora de la investigación, debilita los argumentos sobre los cuales se sostiene el argumento de la *habitación china,* se centra en la estrategia intencional que propone el filósofo Dennett. Antes de sustentar o justificar cualquier argumento basado en dicha estrategia, hay que dejar claro que cualquier cosa puede tomarse como que está haciendo parte de un sistema intencional, pero la forma de explicación difiere dependiendo del sistema, por lo que no se puede utilizar la misma estrategia para todo. De ahí que en un primer momento sea importante mirar otras estrategias que, si bien ofrecen una explicación, la que realmente explica la conducta humana o la de cualquier programa o máquina que interactúe con los seres humanos, es precisamente la estrategia intencional.

Una estrategia que permite ofrecer explicaciones sobre fenómenos naturales, por ejemplo, es la estrategia física. De modo tal que, si se quiere predecir el comportamiento de un sistema, hay que determinar su constitución física en principio; para tal procedimiento se debe recurrir a las leyes de la física y así poder predecir el resultado para cualquier entrada de datos. Aunque Dennett sostiene que esta estrategia es la más más importante, al mismo tiempo es poco practicada, ya que normalmente esta estrategia no se utiliza para hacer explicaciones. Las personas comúnmente no aluden a la física para explicar las reacciones de determinados fenómenos, un ejemplo que plantea el mismo Dennett reafirma lo dicho.

> El químico o el físico puede utilizar esta estrategia en el laboratorio para predecir el comportamiento de materiales exóticos, pero también la cocinera que está en la cocina, puede predecir el efecto de dejar la olla demasiado tiempo sobre el fuego. Esa estrategia no es siempre viable en la práctica, pero que *en principio* siempre funcionará es un dogma de las ciencias físicas. (Dennett, 1998, p. 20)

Aunque la estrategia física podría ser considerada la más importante, y al mismo tiempo la más eficiente para arrojar resultados precisos, hay otro tipo de estrategias que muchos humanos utilizan, cuando se trata de explicar fenómenos. Según Dennett, a veces es más eficaz pasar de la actitud física a lo que él llama la actitud de diseño, donde se desconoce la composición química y física de un objeto; pero se puede predecir su comportamiento teniendo en cuenta el diseño, es decir, cómo está diseñado para comportarse.

"Por ejemplo, la mayoría de quienes usan ordenadores, no tienen la menor idea de qué principios físicos son los responsables del comportamiento altamente fiable y por lo tanto fácil de pronosticar, del ordenador" (Dennett, 1998, p. 20), únicamente conocen cómo funciona y para qué ha sido diseñado el ordenador. Dennett plantea otro ejemplo que parece importante mencionar acá: se sabe la función que cumple un reloj despertador, pero no es del interés conocer y explicar cada detalle de su constitución física, simplemente se asume que sirve para arrojar la hora, programarlo y que suene en un momento determinado.

Si bien, la estrategia del diseño sirve para explicar y predecir el comportamiento de un objeto o fenómeno, en ocasiones es inaccesible acceder a esta; además, puede conducir a explicaciones de carácter mítico o que, no son tan precisas y claras; a diferencia de la estrategia física, la cual predice y explica el comportamiento de un objeto o fenómeno aludiendo a sus propiedades químicas y a las leyes de la física, arrojando una respuesta precisa. No obstante, para explicar las acciones y comportamientos de los seres humanos, la estrategia física no funciona, ya que el contexto en el cual funcionan las explicaciones de las acciones humanas, es un contexto de racionalidad en el que se atribuyen estados mentales, creencias, deseos, intenciones, etc. y la estrategia que mejor permite hacer este tipo de explicaciones, es lo que Dennett llama la estrategia intencional.

Para describir la estrategia intencional hay que tener en cuenta que el objeto sobre el cual se va a predecir se debe considerar como un agente racional, del que se deducen creencias y deseos,

dependiendo de su posición en el mundo. Los seres humanos forjan creencias acerca de todo lo que ven a su alrededor, de ahí que "la exposición a *x*, en otros términos, la confrontación sensorial con *x* durante un lapso adecuado, es la condición *normalmente suficiente* para saber (o tener creencias verdaderas) acerca de *x*" (Dennett, 1998, p. 21), sin embargo, no se aprenden o se recuerdan todas las historias sensoriales que se ofrecen a lo largo de la vida, porque según Dennett, los seres humanos se quedan únicamente con las verdades pertinentes que a las historias sensoriales sirven o que son de su interés. Esto hace parte de una primera regla para atribuir creencias en la estrategia intencional.

Ahora bien, la regla fundamental consiste en atribuir deseos que el sistema debería tener, en ese sentido, se atribuye a los seres humanos una serie deseos básicos, de primer orden, y otros de un orden secundario. Tales como la supervivencia, la alimentación, la procreación, las comodidades, etc. "En forma trivial, tenemos la regla: atribúyanse los deseos de aquellas cosas que un sistema considera buenos para sí" (Dennett, 1998, p. 22). La pregunta que deviene involucra el lenguaje: ¿Qué papel juega el lenguaje en la atribución de deseos y creencias?

El lenguaje permite que se expresen deseos y al mismo tiempo que se atribuyan, "puesto que para obtener lo que se quiere a menudo hay que decirlo" (Dennett, 1998, p. 23). Además, lo que se desea sobrepasa los deseos básicos, dado un contexto sociocultural e histórico y, solo es posible tener acceso a detalles específicos de los deseos si se posee un lenguaje. "Puesto que una vez que uno lo ha declarado, por ser una persona de palabra, uno adquiere interés en satisfacer exactamente ese deseo que declaró y ningún otro" (Dennett, 1998, p. 23).

Pero ¿Cómo funciona realmente esta estrategia? Lo primero que hay que decir, es que esta estrategia, con frecuencia, es utilizada por la gente, como se verá de aquí en adelante. Dennett afirma que dicha estrategia no solo funciona para los seres humanos, sino también para los mamíferos e incluso para las plantas. Por ejemplo, se pueden diseñar y mejorar las trampas de cazas en animales, teniendo en cuenta lo que la criatura sabe o cree sobre diferentes cosas. En otras palabras, esta estrategia permite saber qué le interesa al mamífero dependiendo de su conducta y forma de comportamiento en el medio.

> La estrategia funciona con los pájaros, con los peces, con los reptiles y con los insectos y arañas y hasta con criaturas tan inferiores y poco emprendedoras como las almejas.

> (Cuando una almeja cree que hay algún peligro cerca, no afloja su apretón sobre su concha cerrada hasta que se convence de que ha pasado el peligro). (Dennett, 1998, p. 24)

Adicional a eso, la estrategia intencional también funciona con algunos artefactos. Por ejemplo, en juegos como el ajedrez de computador, se puede decir que este sabe cuáles deben ser las jugadas que le permitirán ganar la partida, en consecuencia, se hace un uso racional a la hora de mover las fichas y por tanto se le aplica un carácter predictivo.

Hasta aquí es claro que tanto los animales, los humanos y los ordenadores hace parte de un sistema intencional, y que la mejor forma de explicar dichos sistemas, se logra a través de la estrategia intencional, ya que esta brinda un poder predictivo que no se puede obtener por ningún otro método, como lo afirma el mismo Dennett. Un científico podría refutar una explicación sobre el comportamiento de un ordenador o un ser humano basado en la estrategia intencional, y argumentar que cualquier explicación reposa sobre las leyes de la física, sin embargo, no se libraría de atribuir la estrategia intencional en su discurso explicativo.

Dennett argumenta y defiende que la estrategia intencional tiene un gran poder predictivo. Plantea unos ejemplos que permiten fortalecer su afirmación, en los que reconoce el carácter explicativo de dicha estrategia: se puede predecir, aunque no con exactitud, el discurso que X político ofrecerá en su emisión, teniendo en cuenta el partido político al cual pertenece y los intereses que defiende. Si se mira este ejemplo y se adapta al contexto propio, fácilmente se pueden predecir las decisiones de los políticos que gobiernan el país, ya que se conocen las ideologías políticas que defienden y sus intereses. Otro ejemplo tiene que ver con los juegos:

> La predicción de los movimientos en un partido de ajedrez; lo que hace del ajedrez un juego interesante es la impredictibilidad de los movimientos del rival, excepto en aquellos casos en que los movimientos son "forzados" —donde *claramente* hay un movimiento mejor— típicamente el menor de los males posibles. Pero esta impredictibilidad se ubica en el contexto cuando uno reconoce que en la situación ajedrecística tipo hay muchísimos movimientos perfectamente legales y por tanto disponibles, pero sólo unos pocos —tal vez media docena— que sean algo recomendable, y de ahí que de acuerdo con la estrategia intencional hay sólo unos pocos movimientos de alta probabilidad. Aun cuando la estrategia intencional no logre distinguir un único movimiento con las mayores probabilidades, puede reducir drásticamente el número de opciones de interés. (Dennett, 1998, p. 25)

Todos los argumentos muestran que la actitud intencional es la mejor forma de explicación para ciertos sistemas, incluso, la objeción que le hace Robert Nozick reafirma el poder de la misma. Nozick planteó el siguiente experimento mental: supongamos que algunos seres de inteligencia superior a la nuestra llegaran al planeta tierra, y que nosotros fuéramos para ellos, lo que son los termostatos para los ingenieros inteligentes. Dicho de otra manera, que no necesitan de la actitud intencional, ni de la actitud de diseño para predecir nuestra conducta, ya que la estrategia física bastaría para ofrecer una explicación microfísica. "Podrían predecir las conductas individuales de los distintos cuerpos en movimiento que observan sin siquiera tratar a ninguno de ellos como sistemas intencionales" (Dennett, 1998, p. 26).

Teniendo en cuenta el experimento planteado por Nozick, ¿se podrían afirmar por ello, que, desde el punto de vista de los seres con inteligencia superior, los seres humanos no tienen creencias? Dennett responde que, si fuera de este modo, concebir a los humanos como seres creyentes no sería algo objetivo, por el contrario, sería algo relativo al observador, siempre que el espectador comparta las limitaciones intelectuales humanas.

> Nuestros marcianos imaginarios podrían predecir el futuro de la raza humana por medio de métodos laplaceanos, pero si no nos vieran también como sistemas intencionales, estarían omitiendo algo perfectamente objetivo: los *modelos* del comportamiento humano que se pueden describir desde la actitud intencional y sólo desde esa actitud, y que sustentan las generalizaciones y las predicciones. (Dennett, 1998, p. 24)

Otro experimento que aclara la imposibilidad de explicar las conductas a partir de un modelo físico es el del marciano y el terrícola. Suponga que un marciano y un terrícola observan la siguiente escena: una señora habla por teléfono y hace una serie de preguntas como ¿Volverás temprano a casa? ¿A qué hora vienes? ¿Vendrás con el jefe a cenar? Finalmente le dice que compre una botella de vino y que conduzca con cuidado. Una vez observan la escena, el terrícola hace la siguiente predicción: vendrá un vehículo con llantas de goma, se detendrá en el camino, traerá una bolsa con una bebida alcohólica, y se bajaran dos sujetos de dicho vehículo; mientras que el marciano hace una predicción adoptando una actitud física, en la que necesita más información de la que sabe, por ejemplo, la velocidad que lleva el vehículo, la composición físico-química del vino, etc.

En pocas palabras, la actitud intencional es inevitable a la hora de explicar la conducta de sistemas inteligentes. Aunque se podría adoptar una actitud física como lo que intenta hacer el marciano en el experimento anterior, sería imposible explicar toda conducta a partir de una actitud física. Además, esos seres con inteligencia superior que se mencionan en ambos experimentos, pueden conversar, teorizar y verse a sí mismos como seres intencionales. En este caso, se puede afirmar abiertamente que tanto las creencias como los deseos no son intrínsecos al sujeto como lo defiende Searle, por el contrario, las creencias y los deseos son relativos al observador, y a la hora de atribuir creencias y deseos, influye el aspecto cultural e histórico de una sociedad.

Dennett asegura que, para ser un verdadero creyente, no se necesita más que un sistema cuya conducta se pueda predecir por medio de la estrategia intencional. De tal forma que, si los ordenadores despliegan ciertas conductas, como en el caso del juego de ajedrez, los humanos como espectadores le podrían atribuir creencias tales como “me quiere ganar la partida”, “quiere ser más listo que yo”, etc. De la misma forma en que unos y otros humanos se atribuyen creencias y deseos entre sí, dependiendo de la forma de comportamiento.

Ahora bien, una pregunta que no solo inquieta al filósofo, sino a los que hacen una lectura filosófica sobre el problema, consiste en saber por qué funciona la estrategia intencional. Una primera respuesta se fundamenta en la evolución, la cual ha diseñado a los seres humanos para ser racionales, para creer y desear, “el hecho de que seamos los productos de un proceso evolutivo largo y exigente garantiza el hecho de que usar la estrategia intencional en nosotros sea una apuesta segura” (Dennett, 1998, p. 31).

Otra explicación a la pregunta, muestra cómo el funcionamiento de la estrategia y del mecanismo, coinciden. Lo que quiere decir que:

> Para cada creencia predictivamente atribuible habrá un estado interno funcionalmente notable del mecanismo, que se pueden descomponer en partes funcionales de casi la misma forma en que la oración que expresa la creencia se puede descomponer en parte, es decir, en palabras o términos. (Dennett, 1998, p. 32)

De tal manera que, las creencias, deseos e intenciones que se les atribuyen a los sistemas inteligentes serán el reflejo de procesos causales.

En conclusión, el argumento que ofrece Searle acerca de que la sintaxis y la semántica son relativas al observador, sirven para debilitar su argumento, ya que se puede afirmar abiertamente que las creencias y los deseos no son intrínsecos, sino relativos al observador, como se ha visto en líneas anteriores. El hecho de que se atribuyan creencias y deseos, está ligado a criterios pragmáticos y criterios que están asociados con la cultura, la historia, el uso de un lenguaje y el manejo de determinados conceptos, los cuales han servido a los humanos para poder interactuar con los demás. En pocas palabras, no hay una cuestión de hecho que permita hacer discriminaciones sobre lo qué es un creyente, porque termina haciendo algo relativo y permeado por con un contexto que podría cambiar. Incluso, teniendo en cuenta el desarrollo tecnológico, podría llegar un momento en la historia, donde se interactuará tanto con las máquinas, que se podría olvidar que son máquinas.

Ahora bien, el argumento que defiende Searle de que únicamente se le puede atribuir pensamiento a algo que tenga los mismos poderes causales que el cerebro, debe ser replanteado. El argumento de las otras mentes, por ejemplo, defiende que para atribuir mente no es necesario la existencia de un cerebro, en ese sentido, se le puede atribuir mente a todo aquello que responda a los estímulos del medio y que pueda interactuar con los humanos.

Referencias bibliográficas

Alarcón, F. (2009). *John Searle y la tesis de la inteligencia Artificial fuerte: críticas y réplicas* (tesis de pregrado). Universidad de Caldas, Manizales, Colombia. Recuperado de https://repositorio.ucaldas.edu.co/handle/ucaldas/4728

Copeland, J. (1996). *Inteligencia artificial.* Madrid: Alianza Editorial.

Cole, D. (2020). The Chinese Room Argument. In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition). Retrieved from https://plato.stanford.edu/archives/spr2020/entries/chinese-room/

Corbí, J. y Prades, J. (2007). El conexionismo y su impacto en la filosofía de la mente. En F. Broncano (ed.), *La mente humana* (pp. 151-174). Madrid: Editorial Trotta.

Changeux, J.P. (1985). *Neuronal Man: The Biology of Mind* (Trad. ing. L. Garey). New York: Pantheon Books.

Dennett, D.C. (1998). *La actitud intencional*. Barcelona: Gedisa.

Descartes, R. (1980). *Discurso del método.* Madrid: Alianza.

Descartes, R. (2007). *Discurso del método. Meditaciones metafísicas* (Trad. M. García Morente). Madrid: Editorial Espasa Calpe.

Hanna, P. (1985). Causal powers and cognition. *Mind*, *94*(373), 53-63.

Martínez-Freire, P. (1996). *La nueva filosofía de la mente*. Barcelona: Gedisa.

Morales, S. (2010). *El problema mente-cuerpo en John Searle* (tesis de maestría). Pontificia Universidad Javeriana, Bogotá, Colombia**.**

Moya-Cañas, P. (2003). El internalismo de los estados mentales en J. Searle. *Acta Philosophica, 12*, 31-62.

Nagel, T. (2000). *Ensayos sobre la vida humana*. México: Fondo de Cultura Económica.

Penrose, R. (1996). *La nueva mente del emperador. En torno a la cibernética, la mente y las leyes de la física*. México D.F.: Fondo de Cultura Económica.

Rapaport, W.J. (1986). Searle's experiments with thought. *Philosophy of Science*, *53*(2), 271-279.

Ruiz Santos, P. (2011). Filosofía de la mente; aportes teóricos y Experimentales a la visión emergentista del vínculo mente-cerebro. *Cuadernos de Neuropsicología, 5*(2), 111-127.

Searle, J.R. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, *3*, 417-458.

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

Searle, J.R. (1994). Mentes, cerebros y programas. En M. Boden (comp.), *Filosofía de la inteligencia artificial* (pp. 82-104). México D.F.: Fondo de la Cultura Económica.

Searle, J.R. (1995). *Mentes, cerebros y ciencia.* Barcelona: Gedisa.

Searle, J.R. (1996). *El redescubrimiento de la mente*. Barcelona: Crítica (Grijalbo Mondadori).

Searle, J.R. (1997). *La construcción de la realidad social.* Barcelona: Paidós Ibérica.

Searle, J.R. (2006). *La mente. Una breve introducción*. Bogotá: Editorial Norma.

Turing, A. (1950). *Maquinaria computacional e Inteligencia* (Trad. C. Cristóbal Fuentes Barassi). Universidad de Chile.

Waldrop, M.M. (1988). Toward a Unified Theory of Cognition, *Science, 241*, 27-29.

Waldrop, M.M. (1988). SOAR: A Unified Theory of Cognition, *Science*, 296-298.

Williams, B. (1996). *Descartes: El proyecto de la investigación pura*. Madrid: Cátedra.