Gerónimo de José Rangel Martínez

Algunos desafío al test de Turing

Introducción. ¿Pueden pensar las máquinas?

Existe un problema en filosofía, a saber, qué es el pensamiento; en torno a esta cuestión, hay otras preguntas, tales como qué se requiere para que algo piense, siendo la pregunta de si las máquinas pueden pensar un caso especial de este último cuestionamiento. Una forma de abordar este problema es por medio de la búsqueda de un conjunto de condiciones necesarias y suficientes que, si son reunidas por algo, entonces estaremos justificados para responder afirmativamente a la pregunta de si esa cosa en particular piensa.

El problema de si una máquina puede pensar no es nuevo, pues Descartes ya había tratado esta cuestión en su *Discurso del método*, donde dio una respuesta negativa a la pregunta. Él nos dice que ni siquiera una máquina que imite en su totalidad nuestra conducta podría pasar como una cosa pensante por dos razones: la primera es que una máquina no podría, bajo ninguna circunstancia, dominar un lenguaje natural tal como los seres humanos lo hacen, esto es, usándolo en diversas situaciones de manera adecuada y efectiva; y la segunda es que, a pesar de que pueda haber máquinas que hagan cosas de manera muy similar a como nosotros las hacemos, e incluso mejor, no podrían hacer todo el conjunto de cosas que un ser humano puede hacer, pues ellas necesitan un mecanismo especial para realizar cada cosa, y es imposible que una máquina tenga tantos y tan diferentes mecanismos.[1] Así, Descartes nos dice que tales máquinas no pueden existir.

Alan Turing trescientos años después, se preguntó si una máquina puede pensar y qué deben ser capaces de hacer estas máquinas para tener pensamiento. Turing y Descartes partieron de puntos

1 *Cfr*. Descartes, René, *Discurso del método*, Juan Carlos García Borrón (editor y traductor), Bruguera, Barcelona, 1968, pp. 159-169.

diferentes para responder la pregunta; el segundo creía en la existencia de dos substancias diferentes, mientras que el primero opta por un enfoque monista.

En su artículo de 1950 titulado *Maquinaria computacional e inteligencia*, Turing ofrece una prueba para determinar si una máquina es inteligente o no. En un primer momento se plantea la pregunta "¿puede una máquina pensar?", pero la descarta porque considera que carece de sentido, pues nos es imposible saber si de hecho la máquina tiene algún pensamiento tal como nosotros lo hacemos, ya que para saber eso tendríamos que ser la máquina misma. Así, plantea una segunda pregunta que considera más adecuada, a saber, "¿puede una máquina imitar a un ser humano?", y para esto se vale de algo llamado "juego de la imitación".

El juego de la imitación consta de tres participantes, donde el participante A y B son de diferente sexo. Los tres sujetos están en habitaciones separadas por lo que la única forma para comunicarse que tienen es por medio de hojas mecanografiadas. Así, el participante C hará una serie de preguntas a ambos para lograr identificar correctamente el sexo de ambos jugadores, pero se debe enfrentar a la dificultad de que el participante A le dará respuestas incorrectas para evitar que logre su cometido; por otro lado, el participante B contestará honestamente a todas sus preguntas. El jugador C ganará si logra hacer la correcta identificación. Sin embargo, Turing hace una modificación substancial a este juego. Propone que en lugar de que el participante A sea un ser humano, supongamos que de género masculino, sea una máquina, y más precisamente una computadora digital. Así, en este primer planteamiento del test tenemos a dos seres humanos y a una máquina, que parece tener la tarea de hacer creer al interrogador que es un hombre. Pero esta primera caracterización del test tiene algunos problemas. Parece ser que, debido a que el interrogador no tiene idea de que hay una máquina haciéndose pasar por un ser humano en el juego, siempre hará una identificación incorrecta, pues sólo considerará si el jugador A es hombre o mujer, no si cabe la posibilidad de que sea una máquina. Luego del planteamiento del test, al parecer Turing da por sentado que el interrogador sabe que el jugador A posiblemente es una máquina.

Después de la publicación del artículo de Turing, la literatura sobre inteligencia artificial proliferó, y se idearon versiones sim-

plificadas del test, que pretendían eliminar ambigüedades como la ya mencionada. Lo que conocemos como test de Turing es una de esas versiones. En esta prueba ya no hay tres participantes, sino solamente dos, a saber, interrogador e interrogado. Igualmente, el interrogador está consciente de que el interrogado puede ser una máquina o un humano.

El test de Turing pretende dar condiciones necesarias y suficientes para determinar si una máquina puede pensar, al igual que cualquier otra cosa que lo realice exitosamente. La condición para que la máquina apruebe el test es que logre, a partir solamente de su conducta lingüística, engañar al interrogador, esto es, hacerlo creer que es un ser humano. Así, la máquina debe imitar la conducta lingüística de un hablante nativo de algún idioma en particular.

Podemos identificar dos afirmaciones diferentes en el test: la del test de Turing y la de la máquina pensante. La primera nos dice que cualquier máquina que lo apruebe puede pensar, sin importar qué tipo de máquina sea y si está programada para simplemente pasar el test; la segunda dice que solamente lo puede pasar una máquina **apropiadamente** programada, pero aquí el problema es qué significa que una máquina esté apropiadamente programada.

A pesar de todo, persiste la duda de qué tipo de máquina es la que se puede poner a prueba en dicho test. La clase de máquina que Turing seleccionó es lo que hoy conocemos como computadora digital, que es aplicación de una máquina universal de Turing. Este tipo de máquinas es capaz de realizar una gran cantidad de tareas gracias a que puede ejecutar una gran diversidad de programas diferentes y a su capacidad de almacenamiento de información[2]. Estas máquinas cuentan con una cinta, finita o infinita, que está dividida en cuadros, y un dispositivo con el que puede escribir símbolos en los cuadros de la cinta para luego leerlos; tiene "estados de máquina", que son especificaciones sobre qué debe hacer, y cuenta con la capacidad de mover la cinta hacia la derecha o hacia la izquierda un cuadro a la vez. Así, una máquina de Turing puede hacer solamente cuatro cosas**:** **1**) puede mover la cinta hacia la izquierda o hacia la derecha un cuadro a la vez; **2**) puede "leer" un símbolo de la cinta a la vez; **3**) puede escribir un símbolo en un cuadro a la vez, ya sea en un cuadro vacío o sobre un símbolo ya

2 El tipo de tareas que pueden realizar son aquellas que pueden ser determinadas por un algoritmo, y que además son físicamente posibles.

escrito en un cuadro; y **4**) puede cambiar sus "estados de máquina" por otros especificados previamente.[3] Es importante aclarar que cualquier computación puede ser realizada por una máquina de Turing, siempre y cuando esté bien formulada.[4]

Casos de computadoras que han logrado realizar tareas que considerábamos genuinamente humanas son por lo menos tres. El primero es el de la computadora Deep Blue, que fue creada en 1996 por IBM con el propósito de competir con el campeón mundial de ajedrez Gary Kaspárov. Pero fue derrotada. En 1997 fue creada una segunda versión, que fue llamada Deeper Blue, la cual logró derrotar a Kaspárov. El segundo caso es el de la computadora llamada ELIZA, creada por Joseph Weizenbaum. Fue una de las primeras computadoras diseñadas para procesar el lenguaje natural. La tarea de ELIZA es imitar a un psicoterapeuta que no dice mucho y simplemente se limita a reafirmar lo que el paciente dice. El programa de ELIZA no es muy largo, solamente toma las oraciones en primera persona del paciente, busca palabras clave, cambia la conjugación de los verbos de primera persona a segunda, y eventualmente añade la frase "por favor cuénteme más". También puede repetir oraciones previas y añadir la oración "¿Qué más puedes decirme sobre esto?".[5] Un caso más reciente es el del programa AlphaGo, que fue desarrollado por Google, y derrotó al campeón mundial de go.

Es fundamental tener en mente una distinción importante en el campo de la investigación sobre la inteligencia artificial, me refiero a la diferencia entre inteligencia artificial fuerte e inteligencia artificial débil. El primero es el programa de investigación acerca de si una máquina puede tener pensamiento tal y como los seres humanos, y junto con esto, toda la gama de estados que consideramos como "mentales"; por su parte, el segundo es el proyecto en el

3 *Cfr*. Crane, Tim, *La mente mecánica. Introducción filosófica a mentes, máquinas y representación mental*, Juan Almela (trad.), Fondo de Cultura Económica, México, 2008, pp. 154-157.

4 Una computación es una serie de pasos para realizar una tarea determinada, con la característica de que los pasos descritos son finitos y no ambiguos. La tesis Church-Turing dice que cualquier cosa para la que haya una computación, entonces esa cosa puede ser realizada por una máquina universal de Turing.

5 *Cfr*. Moody, Todd C., *Philosophy and artificial intelligence*, Prentice Hall, Estados Unidos de América, 1993, p. 87.

que solamente se busca emular algunos procesos que consideramos mentales por medio de computadoras, sin esperar que la máquina de hecho tenga pensamiento. No es claro en qué proyecto estaba pensando Turing, pues en su artículo de 1950 claramente dice que podrá haber máquinas capaces de sentir, lo cual lo compromete con la inteligencia artificial fuerte, pero su postura computacional y matemática define al pensamiento como algo similar al cálculo de operaciones matemáticas, cosa que hace creer que abogaba por una inteligencia artificial débil.

El proyecto de la inteligencia artificial fuerte tiene una serie de desafíos, y algunos de los más fuertes son los que plantean los críticos del funcionalismo, el argumento de la habitación china de Searle, y la "naturaleza" de la inteligencia junto con las limitaciones de las máquinas de Dréyfus.

Antes de presentar las críticas al funcionalismo que atacan al test de Turing y al proyecto de la inteligencia artificial fuerte, primero hay que hacer una caracterización breve del funcionalismo. Según el funcionalismo, que algo sea un estado mental depende únicamente de su función y relación con otros estados, para arrojar ciertos *outputs* cuando se suministran los *inputs* correctos. Así, un estado mental es la función de un estado y el conjunto de relaciones que guarda con otros estados, *inputs* y *outpus*, dentro de un sistema cognitivo.

La relación que el funcionalismo tiene con el test de Turing se encuentra en que las computadoras digitales, en tanto que son máquinas universales de Turing, pueden ser descritas como sistemas funcionales, pues en ellas encontramos estados internos, a los que llamamos estados de máquina, y éstos desempeñan funciones en relación con otros estados internos y los *inputs* que se le suministran a la computadora.

El argumento de los *qualia* ausentes constituye una crítica al proyecto fuerte de la inteligencia artificial, pues apunta a una de las intuiciones más fuertes acerca de qué se requiere para tener pensamiento, a saber, estados fenoménicos de primera persona, tales como el dolor. Este argumento nos dice que imaginemos dos sistemas funcionalmente equivalentes, es decir, con los mismos tipos de estados funcionales realizando el mismo trabajo, y supongamos que ambos se encuentran en el estado funcional de "dolor". Ahora bien, uno de esos sistemas es un ser humano tal como nosotros, y el otro es un robot con millones de homúnculos en la cabeza rea-

lizando el trabajo de los estados funcionales. Si ambos se encuentran en el estado funcional de "dolor", y no tenemos problema para atribuirle ese estado al ser humano, entonces no deberíamos tenerlo para atribuírselo al robot, sin embargo está presente la duda de que se encuentre en el mismo estado funcional y que de hecho sienta dolor. Así, surge la duda de que ambos sistemas sean funcionalmente equivalentes, y de que tener los mismos estados funcionales que algo que piensa es suficiente para tener pensamiento.

Un argumento más contra la propuesta de Turing es que podemos imaginar una computadora que tenga en su base de datos todas las posibles oraciones que puedan interpretarse normalmente como miembros de una conversación en un idioma determinado. Dado esto, siempre que ingresamos una oración a la computadora, ésta rápidamente busca en su base de datos la oración que constituiría una respuesta a la oración que le hemos suministrado como *input*. Una computadora que realice algo como esto podría realizar exitosamente el test de Turing, pero si revisamos cómo es que mantiene una conversación, no es muy seguro que sigamos creyendo que piensa.

Otro caso contra la inteligencia artificial fuerte, y más precisamente contra el test de Turing, es el de la habitación china de Searle. Este argumento parte de un experimento mental que nos dice que nos imaginemos dentro de una habitación en la que tenemos dos cosas: una cesta con todos los caracteres del chino y un manual de la sintaxis de dicho idioma. Luego de un rato, por una hendidura en una de las paredes recibimos una serie de hojas con oraciones en chino, y respondemos a ellas con ayuda del manual y los caracteres. Afuera de la habitación hay hablantes nativos del chino que creen que han estado manteniendo una conversación con la habitación, que realmente es una computadora, e igualmente creían que la computadora había pasado el test de Turing, pues en las hojas en las que respondimos a las que ellos introdujeron había oraciones que constituían respuestas a las preguntas que ellos nos hacían. El punto es que nosotros no entendemos una palabra del chino a pesar de haber manipulado correctamente los caracteres con ayuda del manual.

Podemos reformular el argumento de la siguiente manera. Si la inteligencia artificial fuerte es verdadera, entonces hay un programa para todo idioma, que si es ejecutado por cualquier computadora, ésta entenderá dicho idioma. Yo puedo correr un programa para un idioma determinado sin que eso me lleve a entenderlo. Por lo tanto, la inteligencia artificial fuerte es falsa.

Por último, tenemos la crítica de Huber Dreyfus, la cual versa sobre la dificultad existente para programar una computadora de tal manera que imite el comportamiento humano, pues mientras que la computadora sigue una serie de algoritmos efectivos para realizar una tarea, el comportamiento humano no se basa en algún algoritmo, ya que siempre está sujeto a la variabilidad de las situaciones en las que se encuentra. Si queremos traducir el comportamiento humano a un algoritmo, éste debería estar lleno de cláusulas *ceteris paribus* que incrementarían exponencialmente haciendo imposible determinar cada uno de los pasos a seguir.

El test de Turing, visto desde el proyecto fuerte de la inteligencia artificial, ha recibido diversas críticas, y la mayoría de ellas señalan que no reúne condiciones suficientes para que algo, sea una computadora digital u otra cosa, posea pensamiento. Cabe señalar que no solamente la prueba diseñada por Turing ha recibido este tipo de críticas, pues también otras variaciones del test, como el test total de Turing[6] propuesto por Steven Harnad, han fracasado en la tarea de aclararnos qué se requiere para que algo tenga pensamiento.

Aquí me propongo abordar una serie de críticas realizadas a la prueba diseñada por Turing, comenzando por las que él mismo anticipó en su artículo de 1950, para después pasar a críticas hechas con posterioridad, a saber, las dirigidas al funcionalismo. Algunas de las cuales son de Ned Block, y las de Searle y Dreyfus. El objetivo de este artículo es presentar algunas de las dificultades que enfrenta el test, y así dejar sobre la mesa un breve panorama sobre uno de los más importantes problemas dentro del área de la filosofía de la inteligencia artificial.

Las antiguas críticas

Las objeciones hechas al proyecto de Turing que presentaré a continuación son algunas de las que aparecen en el sexto apartado de

6 El test total de Turing postula que una máquina con pensamiento es aquella que además de imitar la conducta lingüística de un hablante nativo de algún idioma, también imita su conducta física, asegurando así que tenga las relaciones *adecuadas* con el mundo. El problema aquí es qué significa "tener relaciones adecuadas con el mundo".

Computing machinery and intelligence, por lo que Turing ofreció ya una respuesta a éstas, sin embargo no es claro si ha sido una respuesta satisfactoria o no.

La primera que abordaré es la objeción teológica, que parte del principio de que existen dos substancias, una material y otra inmaterial. La combinación de ambas substancias es lo que conforma a una persona, siendo la substancia inmaterial lo que hace que pueda pensar. Así, la sola presencia de la substancia material, como en el caso de una máquina, no es suficiente para que dicho artefacto tenga pensamiento alguno. El carácter teológico de esta objeción se encuentra en que la substancia inmaterial comúnmente ha sido identificada como "alma", y el único ser capaz de unir ambas substancias es Dios.

Se ha observado que esta réplica a Turing tiene algunos problemas, en especial con el dualismo substancial y el teísmo, pero hay otra dificultad con la que se encuentra, pues siendo el caso de que el teísmo y el dualismo substancial estén en lo correcto, no es claro por qué razón Dios no podría unir un "alma" a una máquina, o a cualquier otra cosa que no sea un cuerpo humano.

Ahora pasaré a la objeción conocida como "cabezas en la arena". Esta crítica no es contra la afirmación de que las máquinas pueden pensar, o al menos imitar el pensamiento humano, sino que es en contra de que se construyan máquinas que puedan hacerlo. El argumento parte de la suposición de la existencia de máquinas pensantes, y a partir de eso se consideran las posibles consecuencias indeseables de su existencia, como los terribles escenarios planteados en los relatos de ciencia ficción. Éstas no son las únicas consecuencias indeseables de la creación de máquinas pensantes, se pueden encontrar más solamente usando la imaginación, y luego agregarlas como premisas al argumento, sin embargo, esto se desvía de la cuestión, pues el punto crucial aquí es si *en principio* una máquina puede pensar, no qué pasaría si lo hiciera, siendo lo segundo nada más que especulación.

Pasaré ahora a la objeción matemática. Esta crítica parte del teorema de Gödel, en el que, a grandes rasgos, se nos dice que en al menos un sistema formal hay preguntas incontestables, o enunciados que no pueden ser probados por el sistema, siempre y cuando el sistema use un lenguaje en el que haya tales preguntas; así, si una máquina está sujeta a esta restricción por ser un sistema formal, entonces hay preguntas a las que no puede dar respuesta alguna, así como enunciados que no puede probar, y por lo tanto, fallará el test.

Turing observa que esta crítica a que las máquinas puedan pensar, o imitar el pensamiento humano, solamente es relevante en caso de que los seres humanos no estemos sujetos a dicha restricción, es decir, en caso de que no haya preguntas incontestables para nosotros. Sin embargo, esto no puede dejarse pasar por alto. Supongamos que estamos sujetos a la restricción de Gödel, que hay preguntas incontestables para nosotros; en ese caso no tendríamos forma de distinguir entre una máquina y un ser humano. Ahora imaginemos que no estamos sujetos a dicha restricción, entonces las máquinas, dado que hay preguntas incontestables para ellas, fallarían el test, y no podrían pensar. Según esto, parece que es necesario que para poder pensar no se esté sujeto a la restricción de Gödel, pero esto es muy exigente para el test. Es posible que un ser inteligente que está sujeto a dicha restricción falle el test, y no dé base alguna para poder distinguirlo de una máquina.

La siguiente objeción es la de la conciencia.[7] Quienes la sostienen dicen que es imposible que una máquina escriba un poema o una canción, e incluso que pinte una obra de arte, pues no tiene los sentimientos y emociones necesarios para ello, así como tampoco puede sentir ira, tristeza, placer o alegría. En pocas palabras, las máquinas no tienen conciencia. Y según los defensores de esta posición, tener conciencia es un requisito indispensable para pensar.

La respuesta que Turing da a esta objeción es que la única manera que hay para saber si una máquina siente todas esas cosas, si tiene conciencia, es ser la máquina misma. Igualmente, hace notar que lo mismo pasa con los seres humanos, y si este es el criterio para saber si algo piensa, entonces, para saber si otro ser humano puede pensar tendríamos que ser ese ser humano, sin embargo no podemos hacer tal cosa, ni con máquinas ni con personas. Esto lleva a un solipsismo. Para salir del solipsismo al que esta postura conduce, Turing nos sugiere que tomemos la evidencia de la conducta lingüística de las máquinas, al igual que en los seres humanos, y a partir de eso concluyamos que la máquina tiene mente.

Ahora pasaré al argumento de las discapacidades múltiples. Quienes sostienen esto dicen que una máquina no puede hacer una serie de cosas que normalmente se consideran como una parte

7 Con "conciencia" no se refiere a otra cosa más que a los *qualia*, es decir, a aquellas experiencias fenoménicas de la primera persona que se sienten de tal o cual manera.

fundamental de poder pensar, tales como disfrutar una comida, ser gracioso, amable, iracundo, enamorarse, odiar, etcétera. Según Turing, estas aseveraciones se fundamentan en una inducción, pues quienes las defienden nunca antes han visto una máquina que haga alguna de esas cosas, y concluyen que no puede haber tales máquinas. Sin embargo, desde el punto de vista de Turing, a lo que se debe que una máquina no pueda hacer alguna de esas cosas es a su poca capacidad de almacenamiento, por lo que supone que cuando ésta aumente, entonces podrá haber dichas máquinas con tales capacidades.

Otra de las críticas que le hicieron es la llamada objeción de la informalidad de la conducta. Este argumento parte de que no hay un conjunto de reglas que describan la conducta de una persona en toda situación posible, y por otro lado, que hay un conjunto de reglas que describen la conducta de una máquina en toda situación posible. De esto se puede concluir que los seres humanos no son máquinas, y una máquina no es un ser que pueda pensar.

La respuesta de Turing va encaminada a las consecuencias de que haya reglas que determinen el comportamiento de máquinas y de seres humanos. Si el mundo es determinista, entonces hay reglas tanto para hombres como para máquinas. Si el mundo no es determinista, entonces no hay tales reglas para las máquinas ni para los hombres. Así, en ambos casos, no hay una forma de diferenciar una máquina de un ser humano, pues no se puede determinar qué harán en cualquier posible situación.

Las nuevas críticas al test de Turing

Diversas críticas se han hecho al test de Turing luego de su publicación, y en esta parte sólo abordaré tres de ellas, las que realizaron Ned Block, John Searle y Huber Dreyfus. La crítica hecha por Ned Block no ataca directamente al test, sino que se dirige al funcionalismo, pero como el test se vale de éste, la crítica aplica también para él, pues señala que pasar el test no es una condición suficiente para que una máquina sea inteligente. En cambio, la crítica hecha por Searle parte de la definición misma de lo que es un sistema formal, sin importar los compromisos teóricos que el test tenga. Por otro lado, la crítica de Dreyfus apunta hacia la dificultad existente para programar una computadora de tal manera que

logre imitar el comportamiento humano, pues este último, según Dreyfus, es flexible y no puede ser capturado por un conjunto de reglas finitas.

La crítica de Ned Block

Block tiene dos argumentos contra el test de Turing, el primero es el llamado argumento de *blockhead*, y el segundo es el de los *qualia* ausentes. El argumento de *blockhead* parte de una concepción meramente conductista u operacionista del test y señala que no nos provee de condiciones suficientes para determinar si una máquina es inteligente. Por otro lado, el argumento de los *qualia* ausentes se vale de una aproximación funcionalista del test, y apunta a que a pesar de que haya un sistema funcionalmente equivalente a un ser humano, al cual le podamos adjudicar inteligencia, en principio existe la duda de que en realidad tenga estados mentales cualitativos. Primero expondré el argumento de *blockhead* y después el de los *qualia* ausentes.

En su artículo *Psychologism and behaviorism*, Ned Block sugiere que hay por lo menos una clase de máquina que puede pasar el test de Turing, y que a pesar de esto no es inteligente, dado el *software* mediante el cual se somete al test. Para formular la crítica primero hay que aclarar una cuestión preliminar.

> Llame a una cadena de oraciones cuyos miembros pueden ser tecleados por un humano mecanógrafo uno tras otro en una hora o menos, una cadena *tecleable* de oraciones. Considere la clase de todas las cadenas tecleables de oraciones. Dado que el inglés [y español] tiene un número finito de palabras (y en efecto, un número finito de cadenas tecleables de letras), esta clase tiene un gran, pero finito, número de miembros. Considere el subconjunto de esta clase que contiene todas y sólo aquellas cadenas que son naturalmente interpretables como conversaciones […].[8]

Ahora imagine una máquina programada de tal manera, que tiene la lista de las oraciones que pertenecen al subconjunto de

8 Block, Ned, "Psychologism and behaviorism", en *The Philosophical Review*, Vol. 90, Número 1 (Enero, 1981), p. 19. (Traducción mía)

todas y sólo aquellas cadenas tecleables de oraciones que pueden ser naturalmente interpretadas como partes de una conversación. Cuando la máquina es sometida al test de Turing procede de la siguiente manera: primero, el interrogador teclea una oración **A**, luego la máquina busca **A** en su lista de oraciones, y una vez que la encuentra selecciona una oración **B** para dar respuesta a **A**, formando la cadena de oraciones **AB**, después el interrogador teclea una oración **C**, haciendo que la máquina añada **C** a la cadena de oraciones **AB**, y en seguida de esto, selecciona una oración **D** para dar respuesta a la cadena de oraciones **ABC**.[9] El proceso puede repetirse una y otra vez.

Una máquina como la especificada en el párrafo anterior podría realizar exitosamente el test, permitiéndonos adjudicarle inteligencia, pues sus respuestas serían indistinguibles a las de un hablante nativo del idioma en que se realice la prueba. Sin embargo, si conocemos la forma en la que la máquina mantiene una conversación, nos mostraríamos renuentes a adjudicarle inteligencia. De esta manera, Block señala que el criterio de la imitación de la conducta lingüística de la concepción conductista u operacionista del test de Turing no es una condición suficiente para adjudicarle inteligencia a una máquina, pues hay al menos una que cumple con el criterio pero, como él dice, no es más inteligente que un tostador.

Pasaré ahora al argumento de los *qualia* ausentes, el cual aparece en el artículo titulado *Troubles with functionalism*. El argumento que Block presenta en dicho artículo se vale de una aproximación funcionalista al test de Turing, y más precisamente, de la concepción funcionalista de una máquina de Turing. A continuación expondré el argumento.

Parte del supuesto de que todo sistema con estados mentales puede ser descrito por al menos una máquina de Turing, en la que cada uno de sus "estados de máquina" es idéntico a uno de los estados mentales del sistema. Igualmente, la máquina de Turing y el sistema con estados mentales arrojan los mismos *outputs* ante los mismos *inputs*. Ambos son funcionalmente equivalentes.

Ahora imagine que hay un cuerpo idéntico al suyo, y en la cabeza, en lugar de cerebro, o una computadora, hay una enorme cantidad de homúnculos, aproximadamente la misma cantidad de neuronas que tiene un ser humano adulto. Cada uno de los ho-

9 *Cfr.*, *Ibidem* p. 20.

múnculos tiene una simple tarea previamente especificada, que a grandes rasgos consiste en cambiar una tarjeta de "estado de máquina" o "estado interno" del cuerpo, ante la presencia de un *input* en particular, y arrojar un *output* específico; así, hay un homúnculo que ante la presencia de un I13 y la tarjeta de "estado de máquina" R, arroja el O254 y cambia la tarjeta de "estado de máquina" a T. Así, este cuerpo con homúnculos en la cabeza es funcionalmente equivalente a usted, pues trabaja según la máquina de Turing que describe un sistema con estados mentales.

Ahora bien, en el caso del test de Turing, los homúnculos dentro del cuerpo aparentemente humano se las arreglarían para contestar de la misma manera en que lo haría el sistema con estados mentales, pues ante la presencia de una pregunta constituida por una serie de *inputs* determinados y la presencia de un "estado de máquina", arrojarían un *output* específico y cambiarían a otra tarjeta de "estado de máquina", cosas que están previamente especificadas. De esta manera podrían mantener una conversación con el interrogador sin que éste se percate de que no es un ser humano.

Según Block, lo que hace del robot de cabeza homuncular una réplica al test de Turing es que una vez que se tiene conocimiento sobre el funcionamiento del robot, existe la duda de si, en principio, dicho sistema tiene estados mentales cualitativos.[10] Según esta aproximación funcionalista, para todo sistema con estados mentales genuinos hay una máquina de Turing con "estados de máquina", cada uno de los cuales es idéntico a un estado mental del sistema al cual es equivalente la máquina de Turing. Así, si consideramos el estado mental cualitativo Q de dolor, hay un "estado de máquina" S al cual es idéntico. Sin embargo, como hay duda de que el sistema funcional de la máquina de Turing en realidad tenga estados cualitativos, hay duda de que S sea idéntico a Q, y dado esto, hay duda de que una máquina de Turing sea funcionalmente equivalente a un sistema con estados mentales genuinos. Esto es así porque, si bien no tenemos

10 Un estado mental cualitativo, o *qualia*, es el producto de la experiencia inmediata de la primera persona. En palabras de Nagel, es todo aquello "que se siente de una manera", así, hay algo que se siente como ver el color rojo, como tener dolor de muelas, como el olor del café, etc. Igualmente, Nagel sostiene que tener dichos estados mentales es fundamental para tener conciencia, y que un sistema determinado es consciente si ser como dicho sistema se siente de una manera en concreto.

evidencia concluyente de que un robot de cabeza homuncular tenga estados mentales cualitativos, sí tenemos evidencia de que un cuerpo con cerebro, en lugar de homúnculos, los tiene. Asimismo, si apelamos a una inferencia a la mejor explicación sobre el comportamiento del robot de cabeza homuncular, no es que reacciona como lo hace porque de hecho tenga *qualia*, sino que lo hace porque fue construido para que parezca que los tiene.[11]

El argumento de *blockhead* señala que imitar la conducta lingüística de un hablante nativo de algún idioma sin que el interrogador que forma parte del test se percate de que está hablando con una máquina no es una condición suficiente para adscribirle inteligencia a la máquina que está puesta a prueba. Igualmente apunta a que no es suficiente que tenga conducta observable indistinguible a la de un ser humano, sino que también hay que prestar atención a qué es lo que está pasando *dentro* de la máquina, es decir, cuáles son los mecanismos que están produciendo las respuestas y cómo lo están haciendo. Por otro lado, el argumento de los *qualia* ausentes dice que un sistema funcional no puede ser idéntico a un sistema funcional con estados mentales genuinos porque no tiene estados mentales cualitativos.

Si bien la segunda crítica de Block se puede dirigir al test de Turing debido a que éste se vale de una concepción funcionalista de la mente, pues al ser funcionalista hereda los problemas que se han planteado al funcionalismo, también se pueden esbozar otras críticas, que también aplican al funcionalismo, y especialmente a una máquina que se valga de él.

El problema de cómo pasar de la sintaxis a la semántica en un sistema funcional parte del supuesto del funcionalismo que nos dice que la mente es un conjunto de estados funcionales dispuestos de tal manera que el sistema arrojará los *outputs* adecuados ante determinados *inputs*, en virtud del estado funcional en que se encuentre. Así, el trabajo de un sistema funcional puede ser definido solamente con reglas sintácticas para manipular *inputs* y arrojar *outputs*, según los estados funcionales en los que se encuentre. En el caso de las computadoras esto es más sencillo, pues lo que estaría haciendo el trabajo de los estados funcionales serían los estados de máquina, que son especificados en el programa de la máquina. Este proble-

11 *Cfr.*, Block, Ned, "Troubles with functionalism", en *Minnesota Studies in the Philosophy of Science*, Número 9 (1978), pp. 261-325.

ma es motivo de la crítica de Searle al test de Turing, que será presentada a continuación.

La crítica de John Searle

El argumento de la habitación china parte de un experimento mental, con el que Searle ataca el punto de vista de la inteligencia artificial fuerte. La réplica no trata sobre el material del cual están hechas las computadoras, ni la manera en que están programadas, sino que va a una cuestión más fundamental, a saber, los programas que las hacen trabajar de tal o cual manera. Así, el argumento de la habitación china constituye uno de los mayores obstáculos del test de Turing, y de la inteligencia artificial en general.

Antes que nada, Searle señala hacia cuál postura sobre inteligencia artificial va dirigida su crítica, y es contra la inteligencia artificial fuerte, la cual dice que una computadora, o máquina, apropiadamente programada tiene literalmente una mente y es inteligente. Esta postura es diferente a la de la inteligencia artificial débil, la cual dice que una computadora puede simular diversas capacidades mentales sin que ésta tenga mente y sea inteligente. El experimento mental es de la siguiente manera.

Searle nos pide que nos imaginemos a nosotros mismos dentro de una habitación que, si la vemos por fuera, es una computadora que está siendo manipulada por científicos chinos. La habitación tiene dos ranuras, por una de ellas es por donde se reciben los *inputs*, y por la otra se arrojan los *outputs*. Dentro de la habitación tenemos dos cosas, una canasta con todos los caracteres del chino (que es el idioma nativo de los científicos), los cuales nos parecen totalmente carentes de sentido porque no sabemos chino, y un manual con todas las reglas sintácticas de dicho idioma.

Entonces, los científicos nos suministran un *input*, que consiste en una hoja con una oración escrita en su idioma. Nosotros, con ayuda del manual con las reglas sintácticas y los caracteres, así como un largo periodo de tiempo, logramos formular una oración que constituye un *output* y lo arrojamos por la ranura. El mismo proceso se repite una y otra vez, y nosotros, a pesar de tener la ayuda del manual, seguimos sin entender el significado de los caracteres, los *inputs* y los *outputs*. Por su parte, los científicos creen que la computadora ha pasado el test de Turing, pues han mantenido

una conversación con ella por un largo periodo de tiempo, y ésta no ha dado indicio alguno de ser una máquina, pues sus respuestas son tal como serían las de un hablante nativo del chino, cosa que los lleva a adjudicarle mente, inteligencia y comprensión del chino, cuando lo que en realidad pasa es que quien realiza el trabajo, es decir, nosotros, no entendemos absolutamente nada de este idioma.[12]

El argumento contra la inteligencia artificial fuerte que parte del experimento mental puede ser reconstruido de la siguiente manera:

- Si la inteligencia artificial fuerte es verdad, entonces hay un programa para el chino que si es ejecutado por cualquier sistema computacional, dicho sistema entenderá el chino.
- Yo puedo correr un programa para el chino sin que eso me lleve a entender el chino.
- Por lo tanto, la inteligencia artificial fuerte es falsa.

La premisa dos del argumento es la que está siendo apoyada por el experimento mental. Hay otra manera de reconstruir el argumento, la cual hace más justicia a la crítica de Searle, pues toma en cuenta el aspecto fundamental de la cuestión, a saber, que los programas computacionales son meramente sintácticos, y que la semántica no se puede obtener solamente a partir de la sintaxis.

- La sintaxis no es suficiente para la semántica.
- Los programas computacionales son meramente sintácticos.
- La mente tiene contenido semántico y sintáctico.
- Por lo tanto, los programas computacionales no son suficientes para una mente.

En esta reformulación del argumento, la premisa uno es la que está siendo apoyada por el experimento mental, pues en éste se muestra que no podemos comprender un idioma solamente en virtud de conocer todas sus reglas sintácticas. Por su parte, la premi-

12 *Cfr.*, Searle, John, "Minds, brains and programs", en *Behavioral and Brain Sciences*, Volumen 3 (1980), pp. 417-457.

sa tres parte del conocimiento que tenemos sobre nuestra propia mente, pues sabemos que entendemos el significado del lenguaje.

El argumento señala que ningún programa computacional, dado que todos trabajan únicamente mediante reglas sintácticas, puede llegar a hacer que un sistema que lo ejecute tenga una mente y sea inteligente. Según esto, ninguna máquina que cumpla con las condiciones necesarias para pasar el test de Turing tiene mente e inteligencia.

La forma en que esta crítica puede llegar a afectar al funcionalismo es por lo dicho anteriormente: que los sistemas funcionales, y en especial los que son de índole computacional, es decir, que trabajan con base en un algoritmo y un sistema formal, no hacen más que manipular caracteres mediante reglas sintácticas. Imaginemos un programa para llevar la contabilidad de producto en especie de una panadería; en dicho programa podríamos encontrar palabras como *bolillo, bísquet, sema*, etcétera; ahora imaginemos el mismo programa con la misma finalidad, pero en una ferretería, en esta ocasión encontraríamos palabras como *tornillo, clavo, martillo,* etcétera; ahora bien, si un programa de computadora es meramente sintáctico, no hay problema alguno en usar el programa de la panadería, con exactamente el mismo vocabulario, para llevar la contabilidad de la ferretería, siendo que, por ejemplo, la palabra *bolillo* tomará el papel de la palabra *martillo*, sin que esto genere problema alguno. La cuestión reside en que, al igual que en el argumento de Searle, si un sistema funcional como el de una computadora es meramente sintáctico, entonces de dónde proviene el significado que los estados funcionales aparentemente tienen.[13]

La crítica de Dreyfus

Hubert L. Dreyfus es uno de los más notorios críticos del proyecto de la inteligencia artificial fuerte, y gran parte de su trabajo ha sido motivo de discusión en problemas referentes a la cognición corporeizada. La gran parte de sus críticas, o al menos la que expondré aquí, descansa en el cuestionamiento de cuatro puntos que, según él, los investigadores en el campo de la inteligencia artificial han dado por sentado y son motivo de su especial optimismo con res-

13 *Cfr*. Heil, John, *Philosophy of mind: a contemporary introduction*, Routledge, Estados Unidos de América, 1998, pp. 112-114.

pecto a su campo de estudio. Dichos cuestionamiento los realizó en su libro *What computer can´t do*, y posteriormente en *What computers still can´t do*.

La primer suposición de los investigadores de la inteligencia artificial es la que respecta a la biología, que básicamente dice que el cerebro, a nivel neuronal, procesa información tal como lo hace un ordenador digital, en otras palabras, que los impulsos eléctricos del cerebro están determinados por únicamente dos estados, el de encendido y el de apagado.

La segunda suposición trata sobre la psicología humana, y ésta nos dice que la mente puede ser vista y estudiada tal como se vería y estudiaría un artefacto que procesa información siguiendo reglas formales. Así, los datos suministrados por los sentidos son procesados por la mente, tal como un programa para realizar multiplicaciones procesaría los datos que se le suministren.

Hay una tercera suposición, y es de carácter epistemológico. La suposición epistemológica sostiene que todo lo que puede ser entendido, puede ser formalizado y expresado en términos de relaciones lógicas dentro de un sistema formal gobernado por leyes sintácticas. Así, todo el conocimiento puede ser formalizado.

La cuarta y última suposición es de tipo metafísico, la cual dice que, puesto que todo el conocimiento sobre el mundo puede ser formalizado, entonces la realidad puede ser analizada como un conjunto de situaciones independientes unas de otras, y dentro de dicho conjunto están todas las situaciones relevantes para producir comportamiento inteligente.[14]

La crítica de Dreyfus cuestiona estas cuatro suposiciones. Él nos dice que el comportamiento inteligente del ser humano es producto de su relación con el mundo,[15] pues no puede estar disociado

14 *Cfr*. Dreyfus, Hubert, *What computer still can´t do*, The MIT Press, Estados Unidos de América, 1992, p. 156. Para más sobre el tema de la flexibilidad y capacidad de adaptación del comportamiento humano véase Dreyfus, Hubert & Stuart Dreyfus, *Mind over machine: the power of human intuition and expertise in the era of the computer*, Free Press, Estados Unidos de América, 1988.

15 Lo que Dreyfus quiere decir con *relación con el mundo* parte de la filosofía heideggereana, según la cual, el "ser-ahí" [o *Dasein*] se encuentra siempre situado en el mundo, haciendo algo o teniendo contacto con más cosas del mundo. De esta manera, el ser humano, que según Heidegger es partícipe del "ser-ahí", se encuentra siempre situado en el mundo y haciendo algo en él.

de éste; así, para que un ser humano, e incluso cualquier otra cosa, muestre un comportamiento inteligente debe estar en una constante relación con una diversidad de situaciones distintas. Ahora bien, puesto que las situaciones en las cuales se encuentra el ser humano no pueden ser disociadas de éste, tampoco pueden ser analizadas independientemente, y esto también afecta al conocimiento que se pueda tener del mundo, ocasionando que no se pueda representar mediante un sistema formal gobernado por leyes. Lo que hace que un ser humano exhiba inteligencia no es poder capturar su conocimiento mediante leyes lógicas, ni tampoco es que su mente trabaje tal como lo hace un sistema que procesa información siguiendo un sistema formal, sino que lo que hace posible esto es su capacidad de adaptación y flexibilidad de comportamiento en cada situación posible en la que se vea envuelto, haciendo uso de la *intuición* para tomar decisiones, reconocer patrones, mantener conversaciones, etcétera.[16]

En pocas palabras, no hay manera de formalizar todo el conocimiento humano, ni tampoco hay forma de programar a una máquina de tal manera que se comporte exhibiendo una conducta inteligente en toda situación posible, tal como un sujeto humano. Considérese el siguiente ejemplo: un oficial [humano] de policía tiene una ruta programada, y pasa por allí todos los días. Cada día se enfrenta con diversas situaciones, un día puede ser un asalto, una persecución en automóvil, e incluso puede llegar a encontrarse con la situación de ayudar a una mujer con su parto en plena calle. El oficial de policía solamente tiene órdenes de realizar su ruta diaria y detener a quienes infrinjan la ley, pero siempre se puede topar con situaciones que no estaban previstas, y sin embargo, tendrá que actuar. Así, no hay manera de programar a una máquina para que se comporte tal como lo haría el oficial de policía en toda situación, pues en el algoritmo de la máquina habría una gran cantidad de disyuntos increíblemente grandes.

Siguiendo lo anterior, si se crea una máquina que pueda pasar el test de Turing, ésta no tendría pensamiento, pues para eso es necesaria la flexibilidad de conducta de la cual los seres humanos son capaces y las máquinas no. De esta manera, la crítica de Dreyfus apunta a que el criterio de imitación de la conducta lingüística que propone Turing es insuficiente, pues un ser humano, el cual

16 *Cfr*. *Ibidem* pp. 256-271.

presumiblemente posee pensamiento e inteligencia, hace más que mantener una conversación.

Conclusión

Al inicio de este trabajo presenté algunas de las críticas al test, algunas de las que Turing ya se había encargado en el mismo artículo en el que presentó por primera vez el juego de la imitación. Luego de eso expuse una serie de críticas más recientes, algunas de las que atacaban la postura funcionalista del test, que señalaban la imposibilidad de generar contenidos semánticos a partir únicamente de reglas sintácticas, y las limitaciones de las máquinas.

La primera de las dos críticas de Block, a saber, la de *blockhead*, va especialmente contra el test de Turing, ya que indica que imitar la conducta lingüística de un hablante nativo de un idioma en particular no es suficiente para adjudicarle inteligencia a una máquina. La segunda, la de los *qualia* ausentes, ataca al funcionalismo debido a que cuestiona la tesis de que todo estado funcional es idéntico a un estado mental, poniendo en duda también la existencia de sistemas funcionalmente equivalentes, al menos en lo que respecta a estados mentales sobre experiencias fenoménicas.

La crítica de Searle apunta a una cuestión más fundamental que la equivalencia entre sistemas funcionales, y también más adecuada para el proyecto de la inteligencia artificial fuerte. Él señala que es imposible que una máquina sea inteligente debido a que no podemos obtener contenido semántico alguno únicamente mediante la manipulación de caracteres con ayuda de un conjunto de reglas sintácticas, es decir, que correr un programa no es suficiente para que la máquina piense. Así, esta crítica merma desde la raíz la posibilidad de la existencia de máquinas pensantes.

Por su parte, Dreyfus no pone un obstáculo al proyecto de la inteligencia artificial fuerte desde el comienzo, sino que apunta a una serie de limitaciones a las que están sujetas las máquinas, entre las que destaca la incapacidad de éstas para presentar una conducta flexible al igual que los humanos, impidiéndoles actuar de manera inteligente en toda situación posible.