Capítulo 3

¿Hemos respondido la pregunta "puede pensar una máquina"?

Rodrigo Alfonso González Fernández

Resumen

Este trabajo examina si la pregunta "¿puede pensar una máquina?" ha sido respondida de manera satisfactoria. La primera sección, justamente, examina el *dictum* cartesiano según el cual una máquina no puede pensar en principio. La segunda trata sobre una rebelión en contra de Descartes, encabezada por Babbage. A su vez, la tercera describe una segunda rebelión encabezada por Turing. En ambas se examina, primero el lenguaje mentalista/instrumentalista para describir a una máquina programada y segundo, el reemplazo de la pregunta por el Juego de la Imitación. En la cuarta sección sostengo que la evidencia aportada por dicho juego nos devuelve a la pregunta, pese a Turing. Por último, la quinta sección versa sobre cómo la Habitación China de Searle, paradójicamente, apoya el *dictum* cartesiano, y lo hace porque tal experimento mental se basa en una capacidad mental falible: la introspección. Se concluye que la pregunta acerca de si puede pensar una máquina es derivada de otra pregunta filosófica compleja: la de la naturaleza de lo mental y el problema mente-cuerpo.

Palabras clave: Descartes, dictum, máquina, inteligencia, mente.

Considero que el problema mente-cuerpo es un problema totalmente abierto y sumamente confuso.
Saul Kripke, *El Nombrar y la Necesidad*

1. Introducción

La filosofía es una disciplina que nos enfrenta a preguntas difíciles de responder. Preguntas acerca de la naturaleza de la verdad, del conocimiento, del bien, de la belleza, de lo correcto, de qué es la mente, entre muchas otras, son filosóficamente abiertas. Es decir, el preguntar sobre tales tópicos *no* suscita consenso, sino más bien un debate amplio. Sin embargo, esto no es impedimento para no enfrentar los tópicos a los que refieren dichas preguntas. Es importante notar que los investigadores tratan tales problemas, incluso para decir que no tiene sentido hacerlo. Tal como Kripke (1980) sostiene, no hay consenso filosófico con relación a que es la mente, lo cual muestra que estamos ante una pregunta filosófica abierta, e incluso confusa. Desde que Descartes propuso qué era el pensamiento, esencia de la conciencia y de nuestra vida mental, ha habido múltiples reflexiones, desde aquellas que intentan diluir el problema mente-cuerpo, hasta las posturas que sugieren que la mente simplemente no existe. En consecuencia, la pregunta por la naturaleza de lo mental no solo suscita disenso, sino que además se relaciona con otras preguntas filosóficas difíciles y abiertas, como la que interroga por la posibilidad de vida mental en máquinas programadas o computadores.

En este trabajo muestro que la pregunta "¿puede pensar una máquina?" no ha sido respondida, y que sigue, al igual que el problema mente-cuerpo, siendo un problema filosófico abierto. Ciertamente, las dificultades para caracterizar qué es la mente contaminan la mencionada pregunta, al punto de que hacen difícil una respuesta definitiva. Justamente, en la primera sección examino el *dictum* cartesiano según el cual las máquinas no pueden pensar en principio. Las secciones dos y tres se concentran en las rebeliones en contra del *dictum*, con las propuestas de Babbage y de Turing, respectivamente. Aquél sostiene que el lenguaje mentalista para

describir máquinas es un instrumento, mientras que este parece augurar que la mencionada pregunta no suscitará nunca consenso, por lo que debe ser reemplazada por el Juego de la Imitación. La quinta sección analiza de qué forma dicho juego aporta evidencia inductiva no demostrativa, y ello, como se argumenta, nos devuelve a la pregunta que Turing quiere eliminar. Finalmente, la sexta sección se enfoca en la Habitación China y en cómo, pese a Searle, esta no refuta la IA fuerte, sino que da razones para dudar de que el funcionalismo puede ser una aproximación adecuada a la mente. Finalizo este trabajo con una conclusión en la que muestro que las preguntas filosóficas son en esencia abiertas, y ese es justamente el caso de "¿Tiene X mente?" y "¿puede pensar una máquina?", ambas estrechamente relacionadas.

2. El *dictum* cartesiano: la imposibilidad en principio de que máquinas piensen

La mente es un fenómeno complejo, y lo es por su carácter subjetivo e interno. En efecto, a diferencia de otros fenómenos, especialmente los físicos, la mente es accesible a sí misma, por ejemplo, cuando sentimos un dolor. Los dolores duelen para quien los experimenta, así conocemos que tenemos dolor, un fenómeno de suyo complejo para los demás. Pero, no solo la mente es accesible en un sentido epistémico, de cómo conocemos el fenómeno mismo de manera interna. Su *modo de existencia* subjetivo hace que la mente sea un fenómeno único en el mundo natural. Es decir, existe en tanto la experimentamos conscientemente, y ello la hace uno de los fenómenos más complejos, pero a la vez más fascinantes de estudiar. Es claramente un desafío para la filosofía, disciplina que aspira a la verdad y objetividad, incluso en problemas difíciles como el llamado mente-cuerpo.

El examen de la mente fue propuesto por René Descartes en el siglo XVII, cuando, en su intento de refutar a escépticos y ateos, propuso la existencia del *cogito* (González, 2017). Este es ontológicamente diferente de las cosas materiales por no tener extensión, ni ser divisible y limitado, lo cual traerá importantes consecuencias

para la imposibilidad de que las máquinas piensen, según Descartes. Si bien el francés es juzgado como responsable de habernos legado el problema mente-cuerpo, precisamente por la distinción dualista radical entre el cogito y las cosas materiales, no cabe duda de que funda las bases de la filosofía de la mente, al menos en relación con el examen de la naturaleza de lo mental. Nunca un filósofo había examinado qué era la mente, como un fenómeno que parece distinto de lo material. Este pasaje caracteriza qué es el dualismo cartesiano mediante el argumento de la intuición modal:

> En primer lugar, puesto que ya sé que todas las cosas que concibo clara y distintamente pueden ser producidas por Dios tal y como las concibo, basta con poder concebir clara y distintamente una cosa sin otra, para estar seguro de que la una es distinta de la otra, ya que, al menos en virtud de la omnipotencia de Dios, pueden darse separadamente […] Por lo tanto, como sé de cierto que existo y, sin embargo, no advierto que convenga necesariamente a mi naturaleza o esencia otra cosa que ser cosa pensante, concluyo rectamente que mi esencia consiste en ser solo una cosa pensante, o una substancia cuya esencia o naturaleza toda consiste solo en pensar. Y aunque acaso (o mejor, con toda seguridad, como diré enseguida) tengo un cuerpo al que estoy estrechamente unido, con todo, puesto que, por una parte, tengo una idea clara y distinta de mí mismo, en cuanto que soy solo una cosa que piensa –y no extensa–, y, por otra parte tengo una idea distinta del cuerpo, en cuanto que él es solo una cosa no extensa –y no pensante–, es cierto entonces que ese yo (es decir, mi alma, por la cual soy lo que soy), es enteramente distinto de mi cuerpo, y que *puede existir sin él* (Descartes. 1977, pp. 65-66, AT 78 y 79, énfasis mío).

Pero ese no es el final de la historia dualista, la cual sienta las bases del problema mente-cuerpo, o de cómo dos cosas metafísicamente diferentes se relacionan. Luego, en la misma 6ª Meditación Metafísica agrega:

> Hay una gran diferencia entre el espíritu y el cuerpo; pues el cuerpo es siempre divisible por naturaleza, y el espíritu es enteramente indivisible. En efecto: cuando considero mi espíritu, o sea, a mí mismo en cuanto que soy solo una cosa pensante, no puedo dis-

> tinguir en mí partes, sino que me entiendo como una sola cosa, sola y enteriza. Y aunque el espíritu todo parece estar unido al cuerpo todo, sin embargo, cuando se separa de mi cuerpo un pie un brazo o alguna parte, sé que ello no le quita algo a mi espíritu (p. 71, AT 86).

De este modo, Descartes zanja que espíritu y cuerpo son diferentes, separables, y distintos metafísicamente. Tal dualismo, esencia del problema mente-cuerpo, es fundamental para comprender el *dictum* cartesiano en contra de que las máquinas piensen.

En efecto, en su obra *Discurso del* Método (Frondizi, 1994) Descartes ya había hecho un anticipo de tales argumentos, tanto del descubrimiento del cogito como de su naturaleza metafísica. Un poco antes de proponer el argumento del *cogito ergo sum*, y del de la intuición modal arriba expuesto, introdujo el *dictum* acerca de la imposibilidad de que una máquina piense en principio. Su argumento es el siguiente:

> [Una máquina] no podría ordenar las palabras de formas diferentes para responder al significado que se dice en su presencia, como incluso el menos inteligente de los humanos puede hacer […] Incluso, aunque tales máquinas podrían hacer algunas cosas tal como nosotros les hacemos, o de mejor forma, inevitablemente fallarían en otras, lo cual revelaría que no están actuando de acuerdo con su entendimiento, sino por la pura disposición de sus órganos (p. 113).

En concreto, las máquinas son objetos físicos, y si lo son, tienen partes divisibles y limitadas: engranajes, palancas y mecanismos que hacen que su respuesta al ambiente sea también limitada y automática. Una máquina, entonces, solo responde a los estímulos ambientales por su mera disposición material. Esta opera causalmente, y por conllevar automatismo, esto es, por funcionar en términos de causa y efecto, hace que una máquina sea predecible. A mi juicio, algo similar sucede con los signos naturales (e.g. el humo, los anillos concéntricos de un árbol, etc.), los cuales son efectos de causas. Los animales, si son máquinas, responderían a la misma lógica, cuestión que hace a Descartes presa de críticas

por doquier. Ciertamente, si los animales son máquinas, entonces tienen reacciones al ambiente que son limitadas y predecibles, en función de sus órganos, y por tanto solo emiten signos naturales en virtud de la disposición de dichos órganos.

Por el contrario, el ser humano no puede ser solo una máquina, y esto, estimo, lo capacita para tener respuestas al ambiente que son flexibles, en términos de otros signos, los convencionales. Esto vale algunas aclaraciones previas. En primer lugar, el ser humano, por no ser solo máquina, tiene la capacidad de reaccionar de manera flexible al ambiente. Si esto es así, entonces los seres humanos pueden manejar signos convencionales lingüísticos de modo de significar lo mismo de distintas maneras. Por ejemplo:

O1 Juan ama a María
O2 María es amada por Juan
O3 Juan está enamorado de María

En todos estos casos, un ser humano puede enunciar mediante distintos signos convencionales, ordenándolos de distinta manera. Pero, lo más importante es que puede significar lo mismo incluso utilizando distintos signos lingüísticos.

En segundo lugar, es importante notar que el dominio de un lenguaje, dada la cuestión mencionada de los signos convencionales lingüísticos, indica racionalidad. La razón emplea el lenguaje para comunicar ideas y pensamientos, de modo que los mencionados signos son vehículos de estos últimos. El uso de lenguaje, según el francés, es mediado por el uso de la razón, del cogito. En consecuencia, cualquier organismo que use lenguaje es racional e inteligente.

Finalmente, la razón y el entendimiento son causa de las acciones. En el ser humano sucede todo lo contrario a lo que acontece en los animales-máquina. Estos reaccionan automáticamente, por la mera disposición de sus órganos. En cambio, el ser humano usa su entendimiento para actuar, y así su conducta es siempre guiada por *razones*. Entonces, el entendimiento es visto por Descartes como un instrumento universal, que otorga flexibilidad, tanto en

el uso del lenguaje como en la acción. La conducta humana es guiada por razones, mientras que la conducta animal es explicada por causas y efectos, a partir de la mencionada noción de signo natural.

En síntesis, el *dictum* cartesiano es una consecuencia lógica del dualismo del francés, es decir, de dividir el mundo en cosas mentales y cosas físicas, esencia del problema mente-cuerpo. Como las cosas físicas son limitadas, los *outputs* de las máquinas también lo serán, cuestión crucial con relación a la vida mental. Las máquinas, y los animales, piensa el Descartes metafísico, no tienen mente, y dos signos de ello es que no pueden usar lenguaje y sus acciones son automáticas e inflexibles. De esto se sigue que, en principio, no podría haber una máquina que usara el lenguaje de manera genuinamente humana, y que por ello la Inteligencia Artificial parece completamente condenada al fracaso: si bien es posible crear máquinas que empleen signos convencionales lingüísticos, dicho uso es solo una remembranza de cómo opera la razón. El dualismo, entonces, trae consecuencias importantes para el desarrollo de la Inteligencia Artificial futura. Ello porque, si el francés tiene la razón, la respuesta a "¿pueden pensar las máquinas?" sería negativa en principio y, más aún, definitiva, lo que provocará dos rebeliones materialistas.

3. La primera rebelión de la IA contra Descartes: Babbage

El materialismo, una filosofía monista, es radicalmente diferente al dualismo cartesiano. Básicamente, sostiene que hay una sola clase de substancias en el mundo: las cosas materiales. Todo estaría condicionado por la física, y la mente, pese a ser un fenómeno complejo, sería un fenómeno material más. Ello acontecería pese a la experiencia consciente, la subjetividad y su carácter interno. Ninguno de estos aspectos de la mente impediría que sea un fenómeno físico, porque todo es finalmente material. En consecuencia, el monismo materialista postula que la mente es un fenómeno que, dado el mundo material, sería material también, y ello pese a la subjetividad de la experiencia consciente ligada al *cogito* cartesiano.

Babbage (en Swade, 2000) adscribe a la filosofía materialista, en clave decimonónica. Es decir, en el contexto de la época industrial. En efecto, en tal época se creyó que todo podía ser resuelto con máquinas y que, incluso, la mente era un fenómeno material más. Descartes, por tanto, es el enemigo natural de los materialistas y lo es de Babbage también, pese a que sus ideas no refieren directamente al francés. Con todo, los materialistas buscan refutar a Descartes, argumentando, de una forma u otra, que el hecho de que la mente sea más fácil de conocer que el cuerpo no implica que la mente sea un fenómeno diferente. Más aún, para algunos materialistas como Babbage incluso Dios es una substancia material, tesis que se conoce como panteísmo materialista. Así, para Babbage (en Swade, 2000), *todo*, incluso Dios, es material.

Cabe destacar que sus ideas materialistas responden a una necesidad práctica del siglo XIX, época en que se intenta reducir el mundo a números y máquinas. Hay, ciertamente, una cuestión práctica que aquejó a Babbage, y que lo inspiró para crear y diseñar sus máquinas: la de las Diferencias y la Analítica. A diferencia de lo que ocurre actualmente, antaño no había calculadoras y las máquinas eran diseñadas sin medidas estándar. Ello implicaba una serie de problemas técnicos y prácticos, pero el más importante sin duda era cómo calcular y matematizar el mundo *con certeza*, la finalidad misma de las matemáticas. En el siglo XIX solo existían unas tablas de cálculo y mediante estas se hacían los cálculos complejos de navegación, ingeniería, finanzas, entre otras actividades humanas. Los errores eran frecuentes, cuestión que provocaba, además de pérdidas materiales y humanas, *incertidumbre*.

Las tablas de cálculo estaban plagadas de errores humanos. Teniendo presente este problema, Babbage se propuso crear máquinas que erradicaran el error humano de una vez y para siempre. Es decir, su sueño consistió en que el pensamiento matemático fuera mecanizado, una idea en concordancia con el espíritu de la Revolución Industrial. Su plan consistió en mecanizar las cuatro etapas de la confección de las tablas de cálculo: fórmulas, cálculo, lectura de prueba e impresión. De hecho, Babbage con la Máquina de las Diferencias pretendió abarcar dichas etapas, transformando

la fuerza de palancas y engranajes en pensamiento matemático. Su cometido, no obstante, tuvo una serie de problemas que impidieron la construcción de la máquina en su totalidad. Entre ellos destacó la falta de estandarización de herramientas, problemas financieros y desavenencias entre Babbage y el gobierno británico, producto de su carácter difícil.

Pese a tales problemas, introdujo una serie de mejoras en una segunda Máquina de las Diferencias, lo cual sentó las bases para el diseño de la Máquina Analítica. Esta, gracias a su diseño con unidad de procesamiento, control, memoria e inputs, le valió ser considerada el primer computador de la historia. Y ello sucedió a pesar de que el propio Babbage no imaginó el potencial multipropósito de dicha máquina. En efecto, fue Agusta Ada, Lady Lovelace, la que concibió que los engranajes no solo podrían representar números y operaciones, sino toda clase de entidades: notas musicales, posiciones en juegos de ajedrez y damas, y similares. Es por esto que Ada ha sido considerada por la historia como la descubridora del potencial de la Máquina Analítica, y no Babbage.

Sin embargo, el matemático decimonónico sí descubrió algo digno de notar. Gracias a las mejoras de la segunda Máquina de las Diferencias, y el posterior diseño de la Máquina Analítica, introdujo vocabulario mentalista para describir su funcionamiento. En particular, comenzó a utilizar expresiones como "la máquina recuerda", "memoriza", "piensa", etc. (Swade, 2000) Dicha introducción de vocabulario mentalista fue crucial para el nacimiento de una nueva disciplina filosófica: la Filosofía de la Inteligencia Artificial. Ello porque la pregunta principal que atraviesa a esta es si las máquinas realmente piensan, contra Descartes, o si el uso del vocabulario mentalista es puramente instrumental. Si lo es, entonces no es posible considerar que las máquinas, como los computadores, realmente tienen estados mentales. La atribución de estos solo sería una movida para comprender mejor su funcionamiento. Pero, ¿qué postura tuvo Babbage al respecto?

Su tratamiento del problema filosófico-cartesiano es el siguiente, según Swade (2000): Babbage habla de que al motor [Analíti-

co] "se le puede enseñar a prever". En otro lugar habla de que el motor "conoce". Estaba claro de que usar esta manera de referirse a él era apropiada, y evidentemente sintió que *antropomorfizar* los mecanismos requería justificación o excusa: "La analogía entre estos actos y las operaciones de la mente casi me forzó al uso *figurativo* de estos términos. Fueron adecuados y expresivos y prefiero usarlos en vez de sustituirlos por largos circunloquios" (Babbage en Swade, 2000, pp. 103-104, énfasis mío).

De esta forma, la primera rebelión anti cartesiana busca, mediante el reemplazo del pensamiento matemático por engranajes y mecanismos, sustituir la pregunta "¿puede pensar una máquina?" por un lenguaje que, aunque mentalista, refleje una postura instrumentalista. Tal postura no desecha la pregunta del todo, sino que concentra esfuerzos en contestarla, algo muy similar a lo que busca Alan Turing con su famoso y controvertido Juego de la Imitación.

4. La segunda rebelión de la IA contra Descartes: Alan Turing

Luego de los desarrollos de Turing a propósito del *Entscheidungsproblem* de David Hilbert, y de la pregunta por el potencial de los algoritmos (Turing, 1936), éste concibió un juego que pudiera *reemplazar* la pregunta "¿puede pensar una máquina?" En concreto, con esa iniciativa Turing buscó evitar definir conceptos como "máquina" e "inteligencia", toda vez que no solo llevan a una discusión filosófica sobre el uso común de estos, sino que además al estudio de su significado con base en encuestas tipo *Gallup*. En particular, Turing sostiene lo siguiente en 1950:

> Si la exploración del significado de términos como "máquina" y "piensa" se debe efectuar a partir del análisis de cómo estos se usan regularmente, es difícil evitar la conclusión de que el significado y la respuesta a la pregunta, ¿pueden pensar las máquinas?, debe encontrarse a través de una investigación estadística similar a una encuesta Gallup (p. 40).

Mediante dichas encuestas, piensa Turing, se podría establecer el uso más frecuente de los mencionados términos. Sin embargo, esto sería claramente insatisfactorio para dar fundamento a la Inteligencia Artificial. En efecto, determinar qué piensa la gente comúnmente acerca del significado del término "pensar", por ejemplo, no lleva a un cese de la discusión. Esto es, no lleva a consenso alguno, como sí lo haría la evidencia empírica recabada por un test. Por tanto, Turing descarta la discusión del significado de términos como "pensar", que parece demasiado cercana a filósofos como Descartes y el problema mente-cuerpo.

Como una manera de evitar tal discusión y polémica, plantea un test basado en un juego, el famoso y controvertido Juego de la Imitación. Turing, de hecho, ya había planteado un juego de similares características antes (1948), y luego lo hizo teniendo un presente una filosofía particular (Turing, 1950). En términos generales dicho juego tiene una inspiración funcionalista, al adscribir al principio de realizabilidad múltiple: en relación con propiedades funcionales, por ejemplo, asociadas a estados mentales, no importa el material en que estas se instancian. Tal como un carburador puede ser de cobre, bronce o aluminio, y puede desempeñar la función de mezclar el oxígeno y el combustible, los estados mentales pueden instanciarse en diferentes materiales: en un sistema basado en carbono o silicio, por ejemplo. Más aún, una tesis fundamental del funcionalismo y del principio mencionado es que las funciones son *separables* de los materiales en que se instancian, de un modo similar a cómo la mente es separable del cuerpo, lo que le ha valido a Turing ser acusado de ser dualista encubiertamente y, por tanto, de sostener una postura que no es auténticamente materialista y anti cartesiana (González, 2011).

Independiente de esas acusaciones, es claro, a pesar de lo que sostiene la tradición, que el Juego de la Imitación es funcionalista. Lo es porque la primera versión consiste en que hay un hombre en una pieza, una mujer en otra pieza, y jueces que hacen rondas de preguntas para determinar quién es el hombre y quién la mujer. El rol del hombre es hacerse pasar por una mujer, respondiendo como si fuera una, mientras que el rol de la mujer es ayudar a los jueces a

descubrir que ella es de sexo femenino. Las preguntas son simples, con rondas de no más de 5 minutos de duración. Por ejemplo, una pregunta típica es: ¿Tiene Ud. el cabello largo? La simplicidad de las preguntas es fundamental porque el computador digital no puede estar en una desventaja obvia frente a los jueces.

Por otra parte, un aspecto fundamental del juego, y que habla de su funcionalismo, es que la determinación del sexo de los participantes no es un tópico casual (González, 2015). Según Turing, en la medida que el hombre se hace pasar por una mujer, puede emular la inteligencia femenina, sin que se requiera de cerebro femenino. Lo mismo, por lo demás, acontecería si la mujer se hiciese pasar por un hombre. Para emular la inteligencia de este, no se requeriría que la mujer tuviera cerebro masculino. En ambos casos, los cerebros masculinos y femeninos están constituidos por los mismos materiales, pero se asocian con distintos tipos de inteligencia. Nótese que los materiales, por ser los mismos, son irrelevantes para la existencia de los dos tipos de inteligencia. Así, Turing sostiene la tesis de que esta es una propiedad funcional independiente de los materiales, y es separable de estos últimos, lo que hace que la postura funcionalista de Turing resulte profundamente anti-biológica. Lo es porque no importa la instanciación de la inteligencia en el cerebro; ella podría ocurrir en cualquier sistema que fuera *funcionalmente* equivalente al cerebro, según el funcionalismo. Más aún, es suficiente pero no necesaria la instanciación de la inteligencia en el cerebro.

Ahora bien, el reemplazo de la pregunta "¿puede pensar una máquina?" ocurre en la segunda versión del juego. Turing pregunta al lector: ¿Qué sucedería si un computador reemplazase al hombre en la primera habitación? Con ello, afirma taxativamente, se lograría el reemplazo de aquella pregunta. Nótese que, de manera similar a Babbage, planteó el reemplazo de la inteligencia por la imitación de esta, porque imitar tal inteligencia es suficiente para tenerla. Es decir, el que una máquina sea capaz de imitar la inteligencia hace que esta sea inteligente también. Como una manera de

remarcar que el Juego de la Imitación no involucra una definición de inteligencia, Turing (en Copeland, 2000) afirma lo siguiente en una entrevista concedida a BBC en 1951:

> No quiero dar una definición de qué es pensar, pero si tuviese que darla, probablemente sería incapaz de expresar nada más acerca de ésta que decir que fue un tipo de zumbido mental [*buzzing*] en mi cabeza. Pero no veo que tengamos que estar de acuerdo en una definición en modo alguno. Lo relevante es tratar de distinguir entre las propiedades de un cerebro o las de un hombre, que queremos discutir, y aquellas que no queremos. Para ponernos en un caso extremo, no estamos interesados en el hecho de que el cerebro tenga la consistencia de la papilla. No queremos decir "esta máquina es muy compleja, luego no es un cerebro y no puede pensar". Me gustaría sugerir *una clase particular de test* que uno pudiese aplicarle a una máquina. Ud. podría querer llamar a este un test para ver si la máquina piensa, *pero sería mucho mejor no formularlo así y caer en la petición de principio*, diciendo que las máquinas que pasen el test serán, por decirlo de algún modo, máquinas grado A. *La idea del test es que la máquina tiene que simular ser un humano a través de responder a las preguntas que se le hacen, y sólo pasará este test si la simulación es suficientemente convincente* (p. 6, énfasis mío).

Esta propuesta es, incluso, más radical de lo que parece en primera instancia, al menos desde el punto de vista de la tesis según la cual la inteligencia no requiere de cerebro. En efecto, dado el funcionalismo de Turing, este piensa que una Máquina de Turing puede imitar el comportamiento de cualquier máquina cuya conducta sea en principio calculable. Si esto es así, y el cerebro es una máquina, un computador digital que imite su comportamiento, también tendrá el *output* de un cerebro y, en consecuencia, se podrá sostener que piensa (vuelvo sobre esto más abajo). Con tales ideas, Turing asume que el Juego de la Imitación es prueba suficiente de que las máquinas programadas piensan, o al menos, de que no hay razones para asumir que no lo hacen. Por esto, y como una manera de resumir lo anterior, la tradición ha formulado el Juego de la Imitación estándar, en que se deja fuera el sexo de los participantes.

La versión simplificada del Test de Turing busca aportar evidencia empírica para evitar negar que las máquinas programadas piensan, o de considerar que esto es un sinsentido. En particular, el juego ahora consiste en un computador programado en una pieza, una persona en otra, y rondas de jueces fuera de ambas. La dinámica es similar a las versiones 1 y 2: los jueces formulan preguntas, el computador responde como si fuera humano, y la persona responde de manera sincera, tal como se puede apreciar en la siguiente figura:

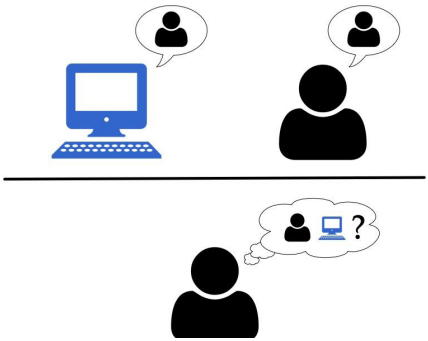

Fig. 1. La versión estándar del Test de Turing, tal como la tradición la ha planteado

Como se puede apreciar, Turing prosigue adhiriendo al funcionalismo, incluso en la versión simplificada estándar del test. No es necesario que una máquina programada o computador tenga el cerebro de un humano para ser inteligente; por el contrario, es suficiente que imite la inteligencia de este para ser inteligente.

En relación con este punto, Turing (1950), incluso, formuló una interesante predicción, con base en el Juego de la Imitación, y a raíz de si puede pensar una máquina:

> Creo que en alrededor de cincuenta años será posible programar un computador, con una capacidad de memoria de 109, para que participe en el Juego de la Imitación tan eficientemente que un interrogador lego no tendrá más de un 70% de probabilidad de hacer la identificación correcta después de cinco minutos de interrogatorio. Creo que la pregunta original, ¿pueden las maquinas

> pensar?, es demasiado absurda para seguir analizándola. No obstante, pienso que a finales de este siglo las ideas de la gente ilustrada y el uso de las palabras habrán cambiado de manera tal que uno será capaz de hablar de máquinas que piensan sin incurrir en contradicción alguna. (p. 49, énfasis mío)

El punto descrito aquí es crucial: el Test de Turing describe un método para la obtención de evidencia *inductiva* que apoye la hipótesis según la cual computadores y máquinas programadas poseen vida mental. Una propuesta, nuevamente, funcionalista y profundamente antibiológica. Sin embargo, el carácter inductivo del test es tan revolucionario como polémico. Y lo es porque, tal como se analiza en la siguiente sección, la evidencia empírica inductiva aportada, que es no demostrativa, no es concluyente en relación con si las máquinas pueden pensar.

5. La evidencia del Test de Turing: ¿fundamenta el reemplazo de la pregunta "puede pensar una máquina"?

Tal como se analizó en la sección previa, Turing reemplaza la pregunta "¿puede pensar una máquina?" por el Juego de la Imitación. Tal reemplazo no considera una definición de inteligencia, sino el aporte de evidencia empírica inductiva que muestre que no tiene sentido negar estados mentales a los computadores digitales o máquinas programadas (Moor, 1976, 1987; Copeland 2000). En este sentido, Turing tampoco pretende que el test sea una condición necesaria de atribución de inteligencia, ni que sea siquiera una condición suficiente. Lo primero, porque no pasar el test claramente no muestra que un sistema no sea inteligente, e.g. los animales. Mientras que lo segundo, la condición suficiente, no se establece claramente en la medida que no hay una definición de inteligencia involucrada (tal como la cita de arriba muestra).

Pero una cuestión que conviene analizar es si el tipo de evidencia proporcionada por el Juego de la Imitación es concluyente con respecto a si una máquina piensa. Parece el caso que Turing confunde dos cosas cuando afirma que el test puede reemplazar

la pregunta de si las máquinas programadas pueden pensar. En efecto, tal parece que confunde un problema del ámbito de la epistemología, esto es, si una máquina podría tener la capacidad de convencer a jueces de que existe inteligencia, con un problema ontológico, a saber, cuáles son las condiciones que son necesarias para la existencia de inteligencia (González, 2007). Más aún, una cosa es *saber* acerca de la existencia de otras mentes, lo cual refiere al problema de las otras mentes, y otra cosa diferente es si dichas mentes existen. Turing parece confundir ambas cosas al plantear que la evidencia empírica inductiva es *suficiente* para el reemplazo de la pregunta "¿pueden pensar las máquinas?".

Puesto de otra manera, confunde el saber acerca de propiedades mentales con la existencia de dichas propiedades. Esto es claro especialmente en lo que respecta a la imitación: nadie diría que imitar la inteligencia *es* necesariamente ser inteligente. Piénsese en una especie de mimo cognitivo que imitase la conducta inteligente de otros humanos, pero al hacerlo no tuviese un criterio de decisión al respecto. En estas circunstancias, ante un error humano, el mimo cognitivo imitaría la conducta, y podría convencer a jueces del caso, incluso si no es inteligente. Por lo tanto, la imitación, base de la persuasión de los jueces, no es garantía suficiente para que *exista* auténtica inteligencia. Ciertamente, una cosa es saber acerca de esta propiedad y otra muy distinta es la existencia de la misma, tal como una cosa es la existencia de otras mentes, y otra el saber acerca de ellas.

El Test de Turing reduce la pregunta acerca de si las máquinas piensan a un método de *verificación* que recabe evidencia empírica de que hay inteligencia. Pero ello no es correcto. El verificacionismo tiene problemas a la hora de responder preguntas ontológicas, al abstenerse de hacerlo producto de su fijación en la conducta observable, científicamente testeable. Tal verificación es adecuada a la hora de corroborar, por ejemplo, si la hipótesis acerca de una partícula sub-atómica tiene sentido. Pero, no sucede lo mismo con la mente y los estados mentales. Tal como Putnam (1965) destaca, con el experimento mental de los super super espartanos, es

perfectamente posible que existan estados mentales sin que exista conducta asociada, o viceversa. Por tanto, la conducta inteligente es contingente en relación con la existencia de estados mentales.

Es importante insistir en que la conducta, que es observable de manera pública y manifiesta, no es útil para responder preguntas de carácter metafísico en relación con la mente. Esto es particularmente relevante en el caso de la metafísica de la mente y de los estados mentales, es decir, de cómo estos existen, un problema subsidiario del problema mente-cuerpo. Justamente, muchos detractores del Juego de la Imitación, en vista de la evidencia que recaba, que es puramente observacional, lo han calificado de conductista, criticándolo como un método *passé* para determinar si los computadores programados tienen efectivamente estados mentales (Block, 1990). Dichos críticos han observado que la conducta lingüística no es suficiente para determinar con certeza que los computadores programados efectivamente poseen mente, estados mentales e inteligencia. En consecuencia, la conducta lingüística, que es observable, no es suficiente para responder de manera tajante preguntas metafísicas acerca de la mente; por ejemplo, si las máquinas programadas tienen estados mentales como los nuestros.

Imitar la existencia de estados mentales no es suficiente para aseverar que dichos estados han sido *replicados*. Claramente no es lo mismo *imitar* la propiedad F que replicar esta: piénsese en la imitación de las condiciones de un huracán y los efectos del huracán mismo. Si alguien afirmara que la imitación de las condiciones es suficiente para tener un huracán, estaría confundiendo el proceso mediante el cual la simulación de este convence a observadores acerca de sus propiedades con la existencia de estas, esto es, con un proceso causal, objetivo e independiente de los observadores. La cuestión de la imitación *versus* la replicación ha sido debatida en filosofía de la mente, especialmente en relación con el experimento mental de la Habitación China de John Searle (como se examinará en la siguiente sección). Turing (en Copeland, 2000), cae, justamente, en este error en una entrevista concedida a BBC en 1951:

> Para lograr que nuestro computador imite a una máquina sólo es necesario programarlo para que calcule lo que la máquina en cuestión haría bajo ciertas circunstancias […] Ahora bien, si una máquina en particular puede describirse como un cerebro, tenemos que solamente programar nuestro computador digital para imitarlo *y también será un cerebro*. Si se acepta que los cerebros reales, descubiertos en animales, y en especial en el hombre, son una clase de máquina, se seguirá entonces que nuestro computador digital, debidamente programado, *se comportará como un cerebro*. Este argumento presupone una idea que puede ser razonablemente cuestionada […] que esta máquina debiera ser de una naturaleza cuya conducta sea en principio predecible mediante cálculo […] Nuestro problema es, entonces, cómo programar una máquina para *imitar al cerebro*, o si lo pudiésemos expresar de una manera más breve y menos rigurosa, *para que piense* (p. 11, énfasis mío).

Pero, hay un elemento que es incluso más importante de discutir, y es el de la evidencia estadística recabada por el Test de Turing. Esta no es suficiente para reemplazar la pregunta de si las máquinas programadas piensan, porque no brinda ninguna clase de respuesta definitiva. Paradójicamente, el propio Turing desecha las encuestas como un método fiable en la determinación del significado de términos como "piensa" y "máquina", y luego propone un juego cuya esencia consiste en brindar evidencia estadística de que los computadores piensan. O al menos, de que no tiene sentido alguno negarles la existencia de estados mentales. Esta paradoja no es trivial, porque conduce de lleno a la pregunta de si las máquinas piensan *de nuevo*. En consecuencia, el método de Turing no es capaz de *reemplazar* definitivamente la pregunta, producto de la evidencia que aporta, puramente observacional, y que ciertamente no es demostrativa en relación con la existencia de estados mentales en máquinas. En efecto, la evidencia observacional no es nunca definitiva en relación con las preguntas que emanan de la metafísica de la mente.

Una segunda crítica que se le hace al test se liga con lo anterior. John Searle (1980) plantea un experimento mental que muestra que, incluso si el test convenciera al 100% de los jueces, ello no

implicaría que una máquina piensa, al menos en los términos como Turing concibe su famoso y controvertido Juego de la Imitación. De este modo, tal juego no solo no ayuda a reemplazar la pregunta, sino que incentiva a intentar una respuesta, desde la arena de la filosofía de la mente. El intento de fundamentar la Inteligencia Artificial por parte de Turing, si bien tiene como objetivo reemplazar la pregunta ¿puede pensar una máquina?, nos devuelve a esta, y lo hace desde una consideración filosóficamente más profunda: la aceptación de que las preguntas sobre la mente no se resuelven con ayuda de la pura evidencia observacional.

6. La segunda ola de críticas al test: la Habitación China

Dado el *dictum* cartesiano, una serie de investigadores en la IA concentraron sus esfuerzos en crear *chatbots* capaces de usar lenguaje natural. Por ejemplo, han creado programas como ELIZA, PARRY, SHRDLU, entre otros. Con excepción de Weizenbaum (1984), creador de ELIZA, los investigadores han sostenido que el manejo de lenguaje natural en conversaciones es conducente a pasar el Test de Turing, signo de inteligencia. Paradójicamente, asumieron el *dictum* cartesiano con el objetivo de refutarlo. En vista de este problema, relacionado con el entendimiento lingüístico, John Searle (1980) distingue entre dos tipos de Inteligencia Artificial, la fuerte y la débil. En particular, propone lo siguiente:

> Encuentro útil distinguir entre la IA "fuerte" y la "débil" (o cautelosa). De acuerdo con la IA débil, el principal valor del computador en el estudio de la mente es que nos brinda una poderosa herramienta. Por ejemplo, nos permite formular hipótesis de modo más riguroso. Pero de acuerdo con la IA fuerte, el computador no es meramente una herramienta en el estudio de la mente; *el computador adecuadamente programado es una mente,* en el sentido de que los computadores con *los programas adecuados pueden decirse que literalmente entienden y tienen otros estados cognitivos,* los programas no son meras herramientas que nos permiten testear explicaciones psicológicas; más bien, los programas son en sí dichas explicaciones (p. 417, énfasis mío).

Tal comentario se fundamenta en un programa específico: SAM, *Script Applier Mechanism,* de Schank y Abelson (1977). Este simula el entendimiento lingüístico de historias, mediante una base de datos que contiene libretos, los cuales son concatenaciones de eventos de manera causal. Por ejemplo, hay un libreto para *Restaurante*, otro para *Trabajo*, y así sucesivamente. Ahora la cuestión interesante es que la aplicación de un libreto a situaciones por parte del computador permitiría a este inferir información que no está explícita en una historia, y ello sería signo de entendimiento lingüístico. Por ejemplo, si un mozo atiende mal mi mesa, a partir de *Restaurante* el computador infiere que no dejaré propina, o que esta será magra. De esta manera, Schank y Abelson simulan el entendimiento lingüístico de historias, lo cual incita a los investigadores de la IA fuerte a pensar que el computador es una mente porque esta entiende historias a la manera de SAM.

Justamente, contra la IA fuerte, y como una manera de *falsar* la creencia de los computadores programados que manejan lenguaje son inteligentes, Searle propone un experimento mental: la Habitación China. Dicho experimento contempla la existencia de una habitación en que hay alguien, hablante de inglés que no habla nada de chino, a quien se le envían ideogramas en este idioma. Gracias a un banco de ideogramas chinos, más un libro de reglas para correlacionar estos, Searle manda hacia afuera de la habitación ideogramas chinos. Sin saberlo, la primera tanda de signos es una historia, la segunda refiere a preguntas y la tercera son respuestas a dichas preguntas. Sin saberlo también, Searle responde de manera perfecta, tal como lo haría un hablante nativo de chino, solo con base en la manipulación de símbolos. Es claro que, bajo estas condiciones, la Habitación China pasaría el Test de Turing, y lo haría convenciendo al 100 % de los jueces de que hay un hablante nativo de chino en la habitación. Con ello, entonces, Searle "falsaría" la IA fuerte (vuelvo sobre esto abajo), en la medida que hace trabajar la mente como esta describe el funcionamiento de la cognición, aunque no habría entendimiento lingüístico genuino.

Hay un fárrago de réplicas a la Habitación China. Teniendo presente el objetivo de este trabajo, no vale la pena ahondar en

todas ellas. Solo me concentraré en dos: la réplica del sistema y la réplica de las otras mentes, pues están directamente relacionadas con la pregunta "¿puede pensar una máquina?" En efecto, la réplica del sistema ataca un punto débil del argumento de Searle: este, con el libro de reglas que es un símil de un programa, es solo una parte del sistema. Luego, incluso si Searle no entiende chino, que es una parte del sistema, no se puede sostener que éste no entiende. La respuesta de Searle es que, si se internalizan en su mente todos los elementos constituyentes del sistema, sigue sin entender. Independiente de la corrección de la respuesta del filósofo norteamericano, es claro que hay un solo elemento que no puede internalizar: él mismo (González, 2012). Este elemento muestra que su experimento mental, al aludir a la introspección, base del análisis de todo experimento mental, tiene una cercanía con el cartesianismo, muy a pesar de lo que Searle piensa. En efecto, la operación llevada a cabo, i.e., la manipulación de símbolos, es evaluada desde la introspección, y ello muestra que solo mediante esta podemos responder la pregunta de si una máquina, en este caso un programa, puede pensar. Como la introspección no es un método totalmente fiable, en tanto puede arrojar resultados experimentales *a priori* falibles no puede concluirse que la Habitación China refuta la IA fuerte de manera definitiva, sino que *solo sienta las bases para dudar del carácter adecuado de tal aproximación* a la mente. Es decir, la Habitación China, un experimento mental *a priori* basado en la introspección, podría, pese a Searle, no refutar la IA fuerte, sino más bien sentar una duda razonable con relación a su verdad.

Hay, además, otro bache en el argumento de Searle: las otras mentes. Dado que la Habitación China se basa en la introspección, no resulta muy claro cómo se puede salvar la objeción de que no es posible saber de la existencia de otras mentes con certeza, cuestión que afectaría a la Habitación China. Si bien Searle responde que lo esencial no es la epistemología acerca de otras mentes, sino la ontología, o las condiciones para la existencia de ellas, queda la duda sembrada, nuevamente, del real alcance de la Habitación China en relación con la refutación de la IA fuerte. Es decir, es claro que sabemos con certeza acerca de nuestras propias mentes, pero no

acerca de la existencia de otras, y ello representa un escollo fundamental para responder adecuadamente si la Habitación China entiende y tiene estados mentales. En consecuencia, no es claro, nuevamente, que dicho experimento *refute* de manera tajante la IA fuerte. Nuevamente, solo puede sostenerse que se *siembra una duda razonable* acerca de la adecuación de esta aproximación a la mente y cognición humana.

La Habitación China no es capaz de brindar certeza con relación a la respuesta a la pregunta "¿puede pensar una máquina?" Ciertamente, tiene la limitación de basarse en la introspección, y aunque se repliquen las condiciones de la IA fuerte, o de otras aproximaciones a la mente a propósito de la mencionada pregunta, no es claro que pueda sostenerse de modo tajante una respuesta definitiva. No hay posibilidad de consenso, entonces, y ello muestra que la Habitación China de Searle es un experimento que no otorga una respuesta definitiva. Por tanto, la pregunta sigue abierta, al igual que lo que ocurre con el famoso y controvertido Test de Turing, todo lo cual es subsidiario del difícil y confuso problema mente-cuerpo.

7. Conclusión

En este trabajo se ha examinado la cuestión de si hemos respondido a la pregunta "¿puede pensar una máquina?" El análisis se efectuó especialmente a la luz del *dictum* cartesiano según el cual es imposible en principio que una máquina piense. Ello ocurriría porque las máquinas son cosas físicas constituidas de partes y engranajes, y por eso son limitadas en relación con los *outputs*, por ejemplo, los lingüísticos. Sin embargo, el *dictum*, subsidiario del problema mente-cuerpo, ha recibido críticas filosóficas importantes. Se examinaron dos: la de Babbage y la de Turing. Mientras que el primero sostiene que la pregunta puede responderse en vista del lenguaje instrumental que se emplea para describir el funcionamiento de una máquina, el segundo sostiene que se puede evitar la pregunta mediante el Juego de la Imitación. Este aportaría evidencia inductiva respecto de que no tiene sentido negar vida mental

a las máquinas programadas o computadores. Luego se analizaron dos respuestas a las aproximaciones de Babbage y Turing: el problema de la evidencia del Juego de la Imitación y el experimento mental de la Habitación China de Searle. Ambas, si bien no son definitivas en cuanto al *dictum*, sí sientan bases para dudar de que hayamos respondido adecuadamente la pregunta "¿puede pensar una máquina?"

Dicha pregunta sigue siendo abierta, y es *filosóficamente* fértil, porque nos lleva a cuestionar la naturaleza de lo mental. Pese a los argumentos de Turing, han pasado más de 50 años a partir de su predicción, y aún no hay consenso en relación con si las máquinas pueden pensar. No existe, así, una respuesta definitiva, lo que indica que la pregunta sigue siendo un incentivo para la investigación en filosofía de la mente. La situación entre partidarios y detractores del *dictum* queda en tablas, y estas, insisto, muestran que la pregunta "¿puede pensar una máquina?" sigue siendo una cuestión filosófica a debatir, un problema que muestra, en todo su esplendor, por qué la filosofía no es una disciplina con respuestas definitivas. La falta de consenso sugiere, además, que la búsqueda filosófica continúa tal como acontece con respecto al problema mente-cuerpo. Hay, pese a Turing y Searle, una relación estrecha entre la naturaleza de lo mental y nuestra posibilidad de responder la pregunta "¿puede pensar una máquina?" Ciertamente, dicha pregunta abre un horizonte de posibilidades y como siempre en filosofía, nos invita a seguir debatiendo y discutiendo en las movedizas y siempre inestables arenas del problema mente-cuerpo.

Referencias bibliográficas

Block, N. (1990). The computer model of the mind. En D.N. Osherson y E.E. Smith (eds.), *Thinking: An Invitation to Cognitive Science*, pp. 247-289. Cambridge, Mass.: MIT Press.

Copeland, B.J. (2000). The Turing Test. En J.H. Moor (ed.), *The Turing Test: The Elusive Standard of Artificial Intelligence*, pp. 1-21. Dordrecht: Kluwer Academic Publishers.

Descartes, R. (1977). *Meditaciones Metafísicas*. Traducción. y notas de Vidal Peña. Madrid: Alfaguara.

Descartes, R. (1994). *Discurso del Método.* Traducción, estudio preliminar y notas de Risieri Frondizi. Madrid: Alianza Editorial.

González, R. (2007). El Test de Turing: dos mitos, un dogma. *Revista de Filosofía Universidad de Chile*, 63: 37-53.

González, R. (2011). Descartes, las Intuiciones Modales y la IA. *Revista Alpha*, 32: 181-198.

González, R. (2012). La pieza china: un experimento mental con sesgo cartesiano. *Revista Chilena de Neuropsicología*, 7(1): 1-6.

González, R. (2015). ¿Importa la determinación del sexo en el Test de Turing? *Revista de Filosofía Aurora*, 27(40): 277-295.

González, R. (2017). La refutación cartesiana del escéptico y del ateo: Tres hitos de su significado y alcance. *Revista Anales del Seminario de Historia de la Filosofía*, 34(1): 85-103.

Kripke, S. (1980). *El nombrar y la necesidad.* México, D.F.: UNAM.

Moor, J.H. (1976). An Analysis of the Turing Test. En S. Shieber (ed.), *The Turing Test: Verbal Behaviour as the Hallmark of Intelligence*, pp. 297-306. Cambridge, Mass.: MIT Press.

Moor, J. H. (1987). Turing Test. En S.C. Shapiro (ed.), *Encyclopedia of Artificial Intelligence*, pp. 1126-1130. New York, Wiley.

Putnam, H. (1965). Brains and behaviour. En J. Heil (ed.), *Philosophy of Mind: A Guide and Anthology*, pp. 96-104. Oxford: OUP.

Schank, R.C., Abelson, R.P. (1977). *Scripts, Plans, Goals, and Understanding.* Hillsdale, N.J.: Erlbaum.

Searle, J. (1980). Minds, Brains and Programs. *The Behavioral and Brain* Sciences, 3(3): 417-424.

Swade, D. (2000). *The Difference Engine: Charles Babbage and the Quest to build the First Computer*. London: Penguin.

Turing, A.M. (1936). On computable numbers, with an application to the Entscheidungsproblem. En M. David (ed.), *The Undecidable.* New York: Raven Press.

Turing, A.M. (1948). Intelligent Machinery. En B. Meltzer y D. Michie (eds.) *Machine Intelligence*, pp. 3-23. Edinburgh: Edinburgh University Press.

Turing, A.M. (1950). Computing intelligence and machinery. En M.A. Boden (ed.), *The Philosophy of Artificial Intelligence*, pp. 40-66. Oxford: OUP.

Weizenbaum, J. (1984). *Computer Power and Human Reason: From Judgement to Calculation.* Harmondsworth: Pelican.

Sobre el autor

Rodrigo Alfonso González Fernández es PhD in Philosophy, por la Katholieke Universiteit Leuven. Actualmente es profesor asistente del Departamento de Filosofía, y Director del Centro de Estudios Cognitivos, de la Facultad de Filosofía y Humanidades, Universidad de Chile. Sus áreas de investigación son la Filosofía de la Mente y de la Inteligencia Artificial y la Ontología Social.