Dos criterios para la presencia de estados mentales: Descartes y Turing

TWO CRITERIA FOR THE EXISTENCE OF MENTAL STATES: DESCARTES AND TURING

Rodrigo González (rodgonfer@gmail.com) Departamento de Filosofía, Universidad de Chile (Santiago, Chile) ORCID: 0000-0001-9693-0541

Abstract

This article examines two criteria for the existence of mental states, namely, Descartes and Turing's. While the former holds that machines can't think in principle, the latter is an advocate of machine intelligence. Despite this, both views seem to be similar in relation to how mental states are judged to exist. Even though this is expected from Descartes' rationalism, it seems surprising in Turing's functionalism. Indeed, no interpreter of the Imitation Game acknowledges that the interrogators can't be replaced and, further, that such interrogators apply an internist criterion to judge whether they are in presence of a mind or a machine. This issue, which involves an internist approach to the mind, makes the Turing Test difficult for a machine. For this reason, I conclude that Turing's test is not only unable to replace the question: can a machine think? In addition, it dramatically shifts the focus of the debate onto the philosophy of mind.

Key words: Descartes, Turing, mind, internism, functionalism.

Resumen

En este artículo examino dos criterios para la existencia de estados mentales, el de Descartes y el de Turing. Mientras que el primero plantea que las máquinas no pueden pensar en principio, el segundo defiende la inteligencia de máquina. Pese a esto, ambos parecen coincidir en que la decisión sobre la presencia de estados mentales es tomada por alguien que juzga internamente la misma. Si bien ello es esperable del racionalismo cartesiano, en el funcionalismo de Turing es sorprendente. En efecto, ningún comentarista de su Juego de la Imitación repara en el hecho de que los interrogadores son una instancia que no parece reemplazable y, además, que dichos interrogadores aplican un criterio internalista para decidir si están frente a una mente o un computador. Tal aproximación internalista a la mente, argumento, representa una dificultad para que una máquina pase el Test de Turing. Por este motivo concluyo que dicho test no solo no reemplaza la pregunta ¿pueden pensar las máquinas?; adicionalmente traslada de modo dramático el foco del debate a la filosofía de la mente.

Palabras clave: Descartes, Turing, mente, internalismo, funcionalismo.

Introducción

A principios de la primera mitad del siglo XVII, Descartes propuso que crear inteligencia de máquina, o crear pensamiento mecanizado, era una proeza imposible de realizar. Para el francés la mente y la inteligencia son, efectivamente, lo mismo, ya que ambas se implican mutuamente. Ser inteligente implica tener estados mentales y tener estos lleva a desempeño conductual inteligente. Junto con esta ecuación,

heredada por parte de la filosofía de la mente contemporánea, el criterio cartesiano para justificar la existencia de otras mentes es controvertido y discutible. Tal criterio se basa en la capacidad de la razón para detectar signos inequívocos de inteligencia, tales como el lenguaje y la acción guiada por el entendimiento. Pero, Descartes lega una complicación extra: se sabe con certeza la existencia de la mente propia, pero no la existencia de estados mentales ajenos. Esto lleva al llamado "problema de las otras mentes", que se caracteriza de manera general, según Hyslop, así: "el problema de las otras mentes es el problema de cómo justificar la casi universal creencia de que otros tienen mente como la nuestra" (2016:1). El problema radica en que hay una asimetría entre nuestra mente, y la justificación de la existencia de estados mentales en los demás, de los cuales solo se observa conducta.

A propósito de dicha conducta, Crane caracteriza más específicamente el problema de las otras mentes en términos de "cómo podemos pasar del conocimiento de la conducta observable de las personas al conocimiento de lo que piensan. Un cierto tipo de escepticismo filosófico dice que no podemos. Este es 'el escepticismo acerca de las otras mentes'. Y el problema que trata es conocido como 'el problema de las otras mentes' […] De acuerdo con este escepticismo, todo lo que realmente sabemos de las otras personas son hechos acerca de su conducta observable. Pero, es posible que la gente *pudiera* comportarse como si no tuvieran mente del todo. Por ejemplo, toda la gente que Ud. ve alrededor suyo podrían ser robots programados por científicos locos para comportarse como si estuvieran conscientes, como si fueran gente pensante: Ud. podría ser la única mente real del todo" (Crane 2003:48).

Un caso que anticipa el problema de las otras mentes es el ejemplo de Descartes de los hombres con capas y sombreros que ve por la ventana. Está directamente relacionado con un proyecto posterior, de la Inteligencia Artificial, tal como Alan Turing originalmente la concibió. En efecto, Descartes afirma que se ven hombres a través de la ventana, pero inmediatamente aclara que las palabras son engañosas: el *juzga* de manera *internalista*, mediante el entendimiento, que son hombres. Lo observable, la conducta que se capta mediante los sentidos, queda en entredicho así, porque debajo de las capas y sombreros podrían haber meros autómatas. Turing, a propósito de la identificación de otras mentes, considera que el carácter observable de estas *se evalúa* por parte de interrogadores: si bien en el Juego de la Imitación se estima que la evidencia lingüística es signo de pensamiento e inteligencia, esta es evaluada de modo *internalista* por dichos interrogadores.

Teniendo presente las similitudes y diferencias entre Descartes y Turing, el presente ensayo tiene como objetivo explorar sus criterios para la existencia de estados mentales. Tal exploración se aborda en cuatro secciones. En la primera se trata con el "ejemplo cartesiano de la cera", el cual antecede el caso de los hombres que se ven a través de la ventana. El caso de la cera es importante, porque muestra:

i) por qué se justifica el conocimiento de la cera mediante la razón;
ii) por qué los sentidos son engañosos;
iii) por qué las palabras son engañosas y desmentidas por la razón.

A propósito de ii y especialmente de iii, erróneamente se cree que se "ve" la cera, un antecedente a considerar en el problema de las otras mentes. En la segunda sección del ensayo justamente trato con como Descartes no cree que se "vean" hombres a través de la ventana. La corrección del juicio, en este caso, se apoya en lo que llamo un criterio internalista de la razón, esto es, en que un sujeto consciente y racionalmente juzga. En la tercera sección analizo de qué forma los interrogadores del Juego de la Imitación también aplican un criterio internalista, basado en lo que ellos juzgan conscientemente. Tal criterio lo aplican a pesar de que el computador está programado para convencerlos de que hay seres

humanos, no máquinas. Finalmente, en la cuarta y última sección, discuto qué consecuencias se siguen de que los interrogadores del Juego de la Imitación tengan un criterio internalista de la mente y de la inteligencia.

1. El ejemplo del trozo de cera: la poca confiabilidad de los sentidos

La certeza del conocimiento es un *leitmotiv* de las Meditaciones Metafísicas. Mediante el orden de razones, Descartes establece que el *cogito* es fundamento del conocimiento cierto. Llega a este de manera sistemática y ordenada, tal como fluyen las mencionadas Meditaciones. Se puede sintetizar el proceso en que el *cogito* establece conocimiento cierto de la siguiente manera: Descartes ha dudado de todo, de los sentidos, de que tiene cuerpo, e incluso considera la posibilidad de que exista un dios engañador que lo engañe respecto de todas las cosas. Ni siquiera los juicios de las matemáticas se salvan de la duda hiperbólica cartesiana, porque el genio maligno podría intervenir cada vez que uno suma 2 + 2 = 4.

Sin embargo, no se puede dudar de la duda hiperbólica, esto es, de que se está dudando; luego, Descartes concluye que piensa y si lo hace, existe. Agrega, además, que este descubrimiento no puede ser una maquinación del dios engañador, porque si lo engañase todo el tiempo, necesitaría pensar. Por lo tanto, nuevamente se corroboraría que piensa y que existe. De esta forma, se establece que no puede dudarse de la existencia del pensamiento. Pero, ¿qué es este? En relación con esta pregunta, Descartes propone que él es una cosa que "duda, que entiende, que afirma, que no quiere, que imagina también, y que siente" (Descartes 1977:26). Todos estos atributos sugieren que el pensar es un concepto *primitivo*, esto es, que no puede definirse en término de otros conceptos, sino más bien por la familiaridad que tiene con los mencionados atributos. Mediante estos se tiene *certeza* de qué es el *cogito*.

Luego de establecer qué es el *cogito*, Descartes se pregunta qué puede saberse con certeza de los objetos materiales. Ahí introduce el ejemplo del trozo de cera. Este tiene una serie de accidentes que hacen suponer erróneamente qué es su esencia. Justamente, en primera instancia el trozo de cera impacta por su color, su sabor, por sonar si se lo golpea, por ser dúctil, por tener figura, entre otras muchas cosas que pueden aprehenderse mediante los sentidos. Es importante destacar que tales cosas *cambian* cuando se somete el trozo de cera al fuego. A causa de este, el color cambia, el olor desaparece, la cera se vuelve líquida, etc. Sin embargo, algo permanece: esto permite decir que se está tratando con cera, y no con otra cosa. Las substancias, a diferencia de los accidentes, subsisten y existen de modo independiente: por eso identifican qué son los objetos. En consecuencia, lo que permanece de la cera se asocia a qué es ella como substancia.

Lo que no cambia de la cera es captado *solo* mediante la razón, no mediante los sentidos. La cera no es la misma luego de someterse al fuego, pero paradójicamente, sigue siendo la misma que Descartes *juzga* como tal. En palabras más simples, las propiedades de la cera captadas externamente, mediante los sentidos, han cambiado, pero dicho objeto sigue siendo lo que es, una substancia: la cera es una cosa física, una *res extensa*. No podría descubrirse esta propiedad de la cera por medio de los sentidos, porque estos solo permiten captar los accidentes de ella. Por este motivo, Descartes afirma que, *finalmente*, la cera, en cuanto *res extensa*, se aprehende mediante el entendimiento y no es "sino solo una inspección del espíritu" (Descartes 1977:29). Esto se relaciona con el idealismo cartesiano: las cosas extensas son finalmente objetos cuyo conocimiento cierto se relaciona con ideas innatas.

Pero, es claro que el análisis del trozo de cera también muestra el racionalismo cartesiano, una aproximación crucial para el problema de las otras mentes. El origen del conocimiento *cierto* radica en la

razón, no en los sentidos. Por este motivo, Descartes constantemente recuerda al lector que los sentidos no son confiables, es decir, no llevan a conocimiento cierto o a *ideas claras y distintas*, como lo pone en la tercera Meditación. En esta propone como regla general "que son verdaderas todas las cosas que concebimos muy clara y distintamente" (Descartes 1977:31). A tales ideas se llega *internamente*, mediante *el ejercicio de la razón*. Los sentidos, por el contrario, llevan a ideas confusas, de las cuales no hay conocimiento *cierto.*

Es importante destacar que los accidentes, aprehendidos mediante los sentidos, se asocian a las palabras ordinarias. Pero, el uso del lenguaje natural es engañoso, porque este sugiere que uno "ve" la cera. También que, por ejemplo, uno "observa" hombres a través de la ventana, en consideración de que solo se ven figuras y ropajes que podrían ocultar meros autómatas. Aunque se ven como hombres, lo relevante es que se los *juzga* como tales, cuestión que tiene directa relación con el criterio cartesiano para la existencia de otras mentes. En efecto, el engaño de las palabras solo puede ser superado mediante el esfuerzo de la razón, de manera *interna*, tal como argumento en la siguiente sección.

2. Palabras engañosas: los hombres que Descartes "ve" por la ventana

En este punto de la discusión conviene recordar el criterio mediante el cual se puede distinguir entre pensamiento e inteligencia, por una parte, e imitación, por otra. Según Descartes, existen dos maneras que reflejan fielmente si se está en presencia de un ser humano con pensamiento e inteligencia, que para el francés son lo mismo. Una es el lenguaje, el cual debe ser juzgado por el uso de signos convencionales lingüísticos, que son vehículos de pensamiento. Una máquina no puede disponer libremente de la manipulación de signos para significar el mismo pensamiento. Por ejemplo, una máquina, por su carácter mecánico, finito y predecible, no es capaz de ordenar las palabras de distintas maneras, tal como lo hace el ser humano.

Su criterio para la existencia de pensamiento e inteligencia se expone del siguiente modo: "Al paso que si hubiera otras [máquinas] semejantes a nuestros cuerpos y que imitasen nuestras acciones cuanto moralmente fuese posible, siempre tendríamos dos medios seguros de reconocer que no por eso eran hombres verdaderos. El primero sería que jamás podrían usar de las palabras ni de otros signos compuestos de ellas como hacemos nosotros para declarar a los demás nuestros pensamientos […] El segundo consiste en que por más que estas máquinas hicieran muchas cosas tan bien o acaso mejor que nosotros, se equivocarían infaliblemente en otras, y así se descubriría que no obraban por conocimiento, sino tan solo por la disposición de sus órganos" (Descartes 1994:112). Justamente, en la tercera sección se discutirá si satisfacer el criterio cartesiano, con las dos condiciones propuestas, implica tener pensamiento e inteligencia.

Es importante notar que en la segunda parte de su argumento Descartes echa mano de la *voluntad libre* de la razón, la cual se opone al mecanicismo para explicitar de qué forma esta y las máquinas se oponen *en principio*. El dualismo fundamenta tal diferencia. La razón tiene un *output* ilimitado mientras que, en el caso de una máquina, compuesta de mecanismos finitos, tendrá un *output* siempre limitado. Más aún, los mecanismos tendrán un *output* que será determinado causalmente, es decir, será afecto a las leyes del mundo físico. Por este motivo, una máquina no puede, en principio, usar lenguaje como un ser humano, pues existe una diferencia de naturaleza entre una cosa constituida por mecanismos finitos, y una *res cogitans*, la cual no tiene límites físicos. Así, la *res cogitans* es ilimitada e incluso inmortal, al ser los límites característicos de lo puramente extenso.

Otro elemento que muestra la diferencia entre el pensamiento y las máquinas es la acción guiada por el entendimiento. Una máquina, que está determinada causalmente y es finita, no puede realizar acciones guiadas por el entendimiento. Por ejemplo, un ser humano lleva a cabo un razonamiento para no caer a un pozo de agua: "si me asomo demasiado, caeré al pozo profundo; luego, no debo asomarme demasiado a este pozo". En cambio, un animal, que para Descartes no es más que una máquina compleja confeccionada por la naturaleza, lleva a cabo acciones automáticas, gatilladas por estímulos, no por la razón. Por ejemplo, si se le pincha una pata al perro, este reaccionará aullando y lo hará solo en virtud de su sistema nervioso. Un ser humano también se quejará ante tal estímulo, pero lo hará porque tiene cuerpo, que es una máquina muy similar a la de los animales. En consecuencia, la razón, entendida esta como instrumento universal, es signo de inteligencia y conducta inteligente, que es guiada por la razón. Esta explica por qué mente, inteligencia y conducta inteligente son co-extensivos para Descartes.

Hay una cuestión crucial con relación al problema de las otras mentes: las palabras resultan engañosas, pues sus significados resultan ambiguos a la luz de los usos. Las palabras desvían la atención de las ideas claras y distintas. En un pasaje posterior al *Discurso* Descartes enfatiza el punto de la siguiente manera: "Finalmente, debido al uso del lenguaje, vinculamos nuestros conceptos a las palabras que usamos para expresarlos; y, entonces, cuando almacenamos los conceptos en nuestra memoria almacenamos simultáneamente las palabras correspondientes. Posteriormente, encontramos más fácil recordar las palabras que las cosas; y, debido a esto, es muy raro que el concepto de la cosa sea tan distinto que podamos separarlo totalmente de los conceptos involucrados en las palabras. Los pensamientos de casi toda la gente tratan más sobre palabras que sobre cosas; y, como resultado, la gente asiente a palabras que no entiende, pensando que una vez que las entendieron, o que las tomaron de otros que sí las entendieron correctamente" (Descartes 1985:220).

Las palabras, entonces, son un impedimento para arribar a ideas claras y distintas, que son constitutivas de conocimiento *cierto*. El común de la gente recuerda las palabras, no los conceptos aludidos por las mismas y referidos a cosas, cuestión que también adquiere importancia para Turing, como se examina más abajo. En efecto, algo similar sucede con las mentes ajenas, las cuales parecen captarse cuando se observa conducta inteligente. ¿Es eso realmente así? Es posible que exista conducta, que existan robots que simulen ser humanos, pero que no tengan mentes. Luego, uno puede caer en el engaño de las palabras, creyendo que *observa* conducta inteligente cuando en realidad no la hay.

El ejemplo del trozo de cera es paradigmático acerca del engaño de las palabras, tal como se analizó en la sección previa. Justamente, en relación con dicho engaño, el filósofo francés recurre al ejemplo de los hombres con capas y sombreros que ve a través de la ventana. Lo hace de la siguiente manera: "Pues aunque estoy considerando ahora esto en mi fuero interno y sin hablar, con todo, vengo a tropezar con las palabras y están a punto de engañarme los términos del lenguaje corriente; pues nosotros decimos que *vemos* la misma cera, si está presente, y no que *pensamos* que es la misma en virtud de tener los mismos color y figura: lo que casi me fuerza a concluir que conozco la cera por la visión de los ojos *y no por la sola inspección del espíritu*. Mas he aquí que desde la ventana veo pasar unos hombres por la calle: y digo que veo hombres, como cuando digo que veo cera; sin embargo, *lo que en realidad veo son sombreros y capas*, que muy bien podrían ocultar meros autómatas, movidos por resortes. Sin embargo, *pienso* que son hombres y de este modo *comprendo mediante la facultad de juzgar*, que reside en mi espíritu *lo que creía ver con los ojos"* (Descartes 1977:29, énfasis mío).

De este modo, al igual que con la cera, Descartes primero asume que "ve" hombre, pero en segunda instancia, luego de examinar su fuero interno, se da cuenta de que son hombres. Nuevamente, queda

claro el carácter racionalista de la filosofía cartesiana, pues los hombres no son identificables mediante los sentidos, sino mediante lo que juzga *internamente* la razón. Alguien podría preguntarse cómo es que se juzgan como tales y, en particular, si no hay participación de los sentidos en la evaluación de si son hombres realmente. En ese momento entra a tallar la cuestión de que los hombres, si solo fuesen captados mediante los sentidos, podrían ser confundidos con autómatas, o para simplificar, con robots. Ello ocurriría porque solo se ven capas y sombreros, y estas prendas podrían ocultar a dichos robots.

Estos no pueden ser distinguidos de los verdaderos hombres solo por medio de los sentidos. Una cuestión que apoya esta idea es que, incluso si los hombres de la ventana fueran pura extensión, al igual que la cera, no serían distinguibles de robots, los cuales también son *res extensa*. Es decir, si uno está en presencia de un robot, o de una máquina, uno no podría distinguirla de hombres solo por lo que se *observa* directamente mediante los sentidos. Ese es precisamente el punto de la analogía de la cera y los hombres que se "ven" por la ventana: no se establece la presencia de hombres en virtud de los sentidos. Cualquier característica podría tenerla un robot, que es una cosa física, y por lo mismo podría ser confundido con un hombre si no se empleara la razón para juzgar qué es. Por lo tanto, que sean hombres o autómatas es algo que se *juzga* mediante una inspección del espíritu, al igual que como acontece con la cera.

Otra cuestión interesante de discutir es *qué* hace juzgar al francés que los hombres que ve por la ventana son hombres y no autómatas o robots. Si se toma en consideración lo examinado hasta aquí, habría que decir que Descartes no confía en lo que solo capta mediante los sentidos. Una hipótesis que se puede aventurar es que debería haber signos que permitan juzgar por qué son hombres y no autómatas. Por ejemplo, si Descartes viera que una dama sonríe y coquetea con uno de los hombres, y luego que este le habla gentilmente, pronunciado palabras y haciendo un gesto cortés, ello contaría a favor de la tesis de que está frente a un hombre, no un robot. Desde la perspectiva cartesiana al menos, los hombres pronuncian palabras y corresponden con gestos cordiales y amables frente a la mirada coqueta de una mujer. Eso es, a no dudarlo, una conducta inteligente que puede juzgarse mediante la razón, puesto que dicha conducta es guiada por el entendimiento.

Todo lo expuesto hasta aquí fundamenta por qué Descartes cree imposible que las máquinas piensen. Es claro que varios pensadores posteriores al francés se opusieron férreamente a la imposibilidad en principio inspirada por su dualismo, idealista y racionalista. Un matemático británico del siglo XX ha sido, indudablemente, uno de los opositores más importantes a tal imposibilidad: Alan Turing. Para este no es menester discutir *conceptualmente* si las máquinas son capaces de inteligencia y pensamiento. Por el contrario, dado que ello llevaría a una suerte de encuesta sobre los usos de palabras como "máquina" e "inteligencia", se debe encontrar una vía alternativa, esto es, un método fiable y certero para reemplazar la compleja pregunta: "¿pueden pensar las máquinas?"

3. Un problema soslayado en el Test de Turing: el criterio de los interrogadores

El Juego de la Imitación es propuesto por Turing como una manera de evitar el lío conceptual que involucra discutir el significado de términos como "máquina", "pensamiento", "inteligencia", etc. Turing parece desconfiar también de cómo se usan las palabras, al punto de que para este un estudio de ellas podría llevar a una encuesta sobre sus usos. Es por esta razón que el Juego de la Imitación, en su conjunto, debe simplemente ponderarse como un intento de *reemplazo* de la pregunta: "¿Pueden pensar las máquinas?" Así es como lo concibe el matemático británico. Podría haber un fundamento no declarado también: tal pregunta podría retrotraer al debate cartesiano y, por tanto, resucitar la propuesta dualista según la cual hay dos clases de substancias.

Sin embargo, vale la pena examinar el Juego de la Imitación brevemente para entender las complicaciones que el caso de los hombres de capa y sombrero anticipa para la Inteligencia Artificial. Si bien el Juego de la Imitación tiene varias etapas, la primera versión puede sintetizarse de la siguiente manera: Turing propone que en una pieza A hay un hombre, en una pieza B una mujer, y afuera hay interrogadores, C, cuyo sexo no es importante. Los interrogadores formulan a A y B rondas de preguntas de 5 minutos con la finalidad de establecer el sexo de los participantes. El objetivo del hombre es *hacerse pasar por una mujer*, respondiendo como si lo fuera. La mujer, B, por el contrario, responde de manera sincera a las preguntas de los interrogadores. Las rondas de preguntas deben responderse de manera escrita para que no dar una ventaja a los interrogadores. Por la misma razón, estos no deben ser expertos, sino gente ordinaria. Pero, ¿por qué Turing introduce el tema del sexo de los participantes y su identificación?

Usualmente se ha desechado la identificación del sexo de los participantes por ser un elemento poco importante en el Juego de la Imitación. El carácter críptico de la primera versión del Juego podría haber ayudado a desestimar la importancia de dicha identificación. No obstante, parece razonable suponer que lo que Turing propone es una concepción funcionalista de los estados mentales, y ello se relaciona justa, aunque no exclusivamente, con la primera versión de su Juego. Lo dice explícitamente, pese a las acusaciones que se le han formulado de haber planteado un test *puramente* conductista. Según esta interpretación del Juego, lo único que importaría es el desempeño lingüístico del hombre dentro de la pieza A. No obstante, parece plausible creer que el funcionalismo de Turing refiere a la imitación de una capacidad, la lingüística femenina, reflejada mediante conducta.

Si Turing es un funcionalista, todas esas críticas soslayan el hecho de que las máquinas tienen estados internos, y de que ellos son clave para generar el *output* con base en el *input* y el programa. Tales estados internos justamente emulan lo que ocurre con los estados mentales, los cuales sirven de puente entre los estímulos y las respuestas a estos. El funcionalismo, a diferencia del conductismo, se basa en la imitación de una capacidad; por ejemplo, en el Juego de la Imitación el hombre imita la capacidad de la mujer para responder como lo haría ella. La imitación de dicha capacidad muestra que no solo la conducta observable importa en el Juego de la Imitación. También es relevante la imitación de la capacidad inteligente femenina, la cual da lugar a un sinnúmero de respuestas atingentes a las preguntas.

Justamente, considero que Turing exhibe simpatía por una visión funcionalista de la mente cuando propone que una ventaja del Juego de la Imitación es que logra separar las capacidades intelectuales de las capacidades físicas del ser humano. Es decir, las capacidades intelectuales no están adscritas a un sistema físico específico, ni *biológico*, porque una función puede ser implementada por distintos sistemas (e.g. eléctricos, hidráulicos, basados en carbono, etc.) Por ejemplo, un corazón artificial puede ser de aluminio o plástico, ya que el material es irrelevante con relación al desempeño de la función. En el caso de la inteligencia femenina, esta no requiere instanciarse en el cerebro femenino, punto por el cual estimo que se introduce la cuestión de la identificación del sexo de los participantes. De este modo, dicha identificación no es tan irrelevante como se piensa: muestra el funcionalismo de Turing, y su aproximación a la mente según la cual la inteligencia es *independiente* de los materiales en los cuales se instancia.

En relación con la distinción entre las capacidades físicas y las intelectuales, Turing manifiesta un compromiso metafísico, además, y este sería incompatible con una concepción puramente conductista de la mente. En efecto, según él “el nuevo problema posee la ventaja de trazar una línea muy clara entre *las capacidades físicas y las intelectuales* del hombre. Ningún ingeniero o químico afirma ser capaz de producir un material que sea indistinguible de la piel humana. Es posible que en algún momento aquello si se podrá hacer, pero aun suponiendo la disponibilidad de esta invención deberíamos sentir que no tiene mucho

sentido en tratar de hacer más humana a una 'máquina pensante' mediante el revestimiento de piel artificial" (Turing 1950:41, énfasis mío).

Subrayo lo de las capacidades en este pasaje, porque confirma el compromiso metafísico del enfoque funcionalista de Turing. Un elemento que puede observarse, como la piel, es irrelevante respecto de si se está en presencia de una máquina inteligente, pensante. Por el contrario, una capacidad, tal como *responder como lo haría una mujer*, es imitable. Precisamente, dada la aproximación funcionalista del matemático británico, para que un hombre se haga pasar por una mujer no es necesario que tenga su corporalidad, o puesto de manera más simple aún, su cerebro. De esta forma, la cuestión de la identificación del sexo de los participantes es atingente para entender la concepción funcionalista y antibiológica de la mente de Turing, una según la cual lo que importa es *cómo funciona la mente* y, en particular, *cómo puede imitarse esta.*

Dado el énfasis funcionalista de Turing en la inteligencia como una capacidad a imitarse, en la segunda versión del Juego de la Imitación se proyecta el reemplazo de un hombre por una máquina programada, o un computador digital. Este se programa para responder las preguntas de los interrogadores *como si fuera una mujer*. Es decir, el computador reemplaza al hombre en la pieza A y es programado para responder como ella, *desempeñando entonces la misma función del hombre*. Lo que se imita, entonces, es la capacidad para responder *como lo haría una mujer*. Gracias a la imitación de dicha capacidad el computador podría lograr un desempeño adecuado para convencer a los interrogadores. Luego, la segunda versión también reafirma el funcionalismo de Turing, en que la conducta sería signo de mente e inteligencia. Esta parece buscar satisfacer el estricto criterio de Descartes descrito en la segunda sección.

Es importante tener en consideración que las máquinas involucradas en el Juego son computadores digitales de capacidades suficientes para realizar la imitación. Es decir, no es válido objetar a las máquinas diciendo que, en cierto momento histórico, estas no tienen capacidades suficientes para imitar a una mujer. Lo que importa es que es *lógicamente posible*, o en palabras simples, que puede concebirse una imitación de tales características, pese a que en la época de Turing fuera *técnicamente imposible* llevarla a cabo. Que algo sea técnicamente imposible no implica que sea lógicamente imposible. Por ejemplo, en la actualidad es técnicamente imposible saber cuántos sistemas solares son similares al sistema solar en Andrómeda. Pero ello no es imposible de verificar de alguna forma, que todavía no se ha desarrollado técnicamente. En consecuencia, las máquinas involucradas son computadores digitales *posibles*, esto es, Turing sostiene que es lógicamente posible que existan máquinas programadas que eventualmente engañen a los interrogadores humanos.

Turing hizo una predicción, a propósito de las máquinas involucradas y de la forma en que se podría sortear la imposibilidad técnica. En el pasaje de la predicción, Turing afirma que "en un periodo de tiempo de 50 años será posible programar computadores, con una capacidad de almacenaje de alrededor de 10^9, para que puedan jugar el juego de la imitación de tal manera que el interrogador promedio no pueda obtener más de un 70% de posibilidades de hacer la identificación acertada luego de cinco minutos de preguntas. Pienso que la pregunta original '¿pueden pensar las máquinas?' es demasiado sin sentido como para merecer discusión. No obstante, creo que cuando lleguemos a finales de siglo, *el uso de las palabras y la opinión educada general habrán cambiado tanto*, que uno podrá ser capaz de hablar de máquinas pensantes sin esperar que exista una contradicción" (Turing 1950:49, énfasis mío).

Tal como señalo arriba, el matemático inglés parece pensar que el uso de las palabras es engañoso cuando se trata de identificar estados mentales, signo de inteligencia. En particular, el uso del lenguaje cambia a

través del tiempo, y lo hará, de acuerdo con la predicción, cuando los computadores digitales tengan capacidades suficientes para pasar el test. Esto supuestamente debería haber ocurrido alrededor del año 2000, es decir, deberían haberse desarrollado máquinas con velocidad y almacenamiento suficientes para imitar a un ser humano, de acuerdo con la versión estándar, simplificada, del Juego de la Imitación.

En dicha versión se asume que A es un computador digital, B es un ser humano, y C son interrogadores humanos. A mi juicio, una desventaja de esta versión es que de alguna forma desdibuja el carácter antibiológico del funcionalismo de Turing. El punto de la conveniencia de la simplificación es polémico, no obstante, porque el propio matemático británico terminó asumiendo la versión estándar luego de 1951, tal como dos entrevistas concedidas a BBC lo confirman. En una de ellas dice que "es probable, por ejemplo, que en el fin del siglo será posible programar a una máquina para responder preguntas de tal forma que será extraordinariamente difícil saber si las respuestas son dadas por un ser humano o una máquina" (Turing 1951:114). El énfasis en la conducta lingüística de seguro ha desorientado a los comentaristas, haciéndoles creer que el Test *solo* es una definición conductista de la inteligencia humana.

Esto podría explicar por qué a Turing se le adjudica una visión externalista de la mente y de la inteligencia. Es decir, una aproximación según la cual lo relevante en el Test es que los computadores causen mediante las respuestas la creencia de que se está en presencia de un ser humano, no de una máquina. Tal como recalco en otro trabajo: "De alguna forma, Turing pone la carga de la prueba sobre aquellos que piensan que es necesario, metafísicamente, replicar la inteligencia para que algo sea inteligente. Para él, en cambio, es *suficiente* desde un punto de vista epistemológico que seamos convencidos de que A es inteligente. En consecuencia, Turing intenta eliminar una discusión metafísica para centrar el debate en la inteligencia de máquina sobre bases epistemológicas, a saber, aquellas en que se reemplaza la pregunta '¿Pueden pensar las máquinas?' con la evidencia empírica aportada por su juego" (González 2015:283).

Sin embargo, hay un aspecto del Test que no ha sido discutido suficientemente, y que se vincula con el problema de la detección de los estados mentales y la inteligencia en los interrogadores del Juego de la Imitación. Claramente existe un problema inadvertido en dichos interrogadores, lo cual no ha sido explorado suficientemente. Los interrogadores, pese a ser no expertos, tienen un *criterio* mediante el que identifican a los participantes en función de lo que *debieran* responder a las preguntas. Una cuestión interesante es que Turing no considera el reemplazo de los interrogadores, o que haya un programa que desempeñe la misma función que los mismos.

Este carácter irremplazable de los interrogadores evidencia que estos tienen un criterio internalista en relación con la mente y la inteligencia. En efecto, los interrogadores *saben* qué respuestas resultan más adecuadas si hay reemplazo de una mujer, o de una persona. Alguien podría argumentar que la cuestión de los interrogadores es anecdótica y que Turing nunca pensó en ellos como elementos a reemplazar en el Juego. Adicionalmente, podría argüirse que la evidencia observacional, de las respuestas dadas frente a las preguntas, es de suyo *suficiente* para concluir si se está o no frente a una máquina. Si bien esto es en parte correcto, no es menos cierto que el criterio para establecer si se está en presencia de hombres, mujeres o máquinas es, al menos respecto de los interrogadores, *internalista*. Ciertamente, los interrogadores, al igual que Descartes, juzgan y saben *conscientemente* qué debiera responderse. Así establecen si están en presencia de mente e inteligencia, o bien de una máquina.

Vale la pena aclarar una vez más la cuestión del internalismo de los interrogadores. Al decir que su criterio para la existencia de mentes es internalista quiero decir que privilegian el juicio interno, consciente, que cada uno de ellos hace. Se evalúa si, por ejemplo, A ha respondido *como lo hubiera hecho una mujer (o*

como una persona), como la capacidad que tiene esta para responder como un cerebro femenino. Por ejemplo, si el computador respondiera a la pregunta ¿qué tipo de pelo tienes?, diciendo "no me gusta depilarme", ello claramente contaría como una respuesta poco adecuada a la pregunta. Enfatizo que los interrogadores *juzgarían*, en función de lo que debiera haberse respondido, que la respuesta es inadecuada. Y para realizar dicho juicio, internamente, los interrogadores deben estar *conscientes* de qué debiera responderse en cada caso.

A propósito de este último punto, estimo con base en lo examinado hasta aquí que el Test de Turing es *difícil* para los computadores digitales puesto que, pese a las precauciones de su autor, es improbable que los humanos *juzguen* que la máquina ha pasado el test de manera efectiva y, sobre todo, que la evidencia recabada es *cierta* e *indiscutible*. La dificultad del Juego de la Imitación, al menos en lo que concierne a las decisiones de los interrogadores, es clara. El criterio que emplean dichos interrogadores es *internalista*: basado en lo que la razón juzga interna y conscientemente. En ambos casos no hay elementos externos que permitan, *por sí solos*, determinar que se está frente a hombres con capas o sombreros, o bien frente a humanos, como acontece en el Juego de la Imitación.

Para poner el punto de una manera más clara todavía: en Descartes y en Turing hay evidencia observable disponible para determinar si hay mente e inteligencia, las capas o sombreros y la conducta lingüística, respectivamente. Sin embargo, dicha evidencia se *juzga* adecuada o inadecuada desde un punto de vista interno, *consciente*. No es de extrañar, entonces, que el Test de Turing sea tan polémico, y que ni siquiera hoy esté claro si alguna máquina programada lo pasó efectivamente. Ciertamente, el internalismo posee una desventaja clara en relación con un test para la inteligencia, a saber, hace presa la decisión de la existencia de otras mentes a una mente individual, la que *juzga* si la evidencia observacional disponible es *cierta*.

Por esta razón la discusión, contrariamente a lo que pronosticó Turing, se trasladó de la arena conceptual de los términos "máquina" e "inteligencia" a su test. Ello explica por qué la pretensión de acallar el debate filosófico claramente no se cumplió. Aunque el nuevo debate no es estrictamente conceptual, ni metafísico, en el sentido de que no vuelve otra vez a la metafísica dualista cartesiana, la evidencia recabada por el test también puede cuestionarse, al punto de que se sigue discutiendo si las máquinas han sido exitosas en el Juego de la Imitación. En relación con el carácter inductivo que tendría el Test considero que, aunque lo tuviese, igualmente daría lugar a una discusión filosófica. En efecto, aportaría evidencia observable respecto de que una máquina programada tiene estados mentales, y es así inteligente. Pero dicha evidencia no sería contundente y definitiva, es decir, no sería considerada *cierta*, incluso si todos los interrogadores fueran engañados.

4. La discusión que suscita el criterio internalista de los interrogadores en el Test de Turing

Uno de los experimentos mentales más significativos para evaluar la efectividad del Test de Turing es la *Habitación China* de John Searle. A pesar de que el argumento es antiguo, y tan controvertido como el propio Juego de la Imitación, hay un elemento de él que resulta interesante destacar. Searle sostiene que "una manera de testear una teoría de la mente es preguntarse a uno como sería [*what it would be like*] si mi mente trabajase bajo los principios que la teoría supone con que la mente trabaja" (1990:68). De esta manera, el desiderátum de Searle lleva a una pregunta sobre qué le ocurriría a uno si mi mente operase de acuerdo con lo que establece la Inteligencia Artificial fuerte. Es decir, qué sucedería, qué se seguiría, si esta teoría fuera efectivamente verdadera.

Tal pregunta lleva a Searle a plantear un experimento mental. Este describe a Searle, un hablante de Inglés que no sabe nada de Chino, en una habitación. Manipula símbolos o ideogramas en Chino, con base en la forma de ellos, y en función de un libro de reglas en Inglés. Pese a que se manipulan los símbolos, y que se puede responder a preguntas sobre historias en Chino, no hay entendimiento lingüístico genuino. De hecho, un conjunto de hablantes nativos de Chino juzgaría, fuera de la pieza, que hay un hablante de Chino dentro de esta. Para evitar cualquier confusión relacionada con la manipulación de símbolos: no hay entendimiento lingüístico en el caso de que *solo* haya manipulación, o sintaxis, porque esta es, según Searle, insuficiente para la semántica. A mi juicio, el hecho de que él sea el *locus* del experimento mental muestra que el cartesianismo está *ciertamente* presente en su *gedankenexperiment*.

Searle sabe *internamente* si hay entendimiento lingüístico o no. Como recalco en otro ensayo: "El punto de vista de la primera persona del agente cognitivo es lo que permite contrastar la manipulación simbólica con el entendimiento genuino, cuestión crucial al ejecutar el experimento. Solo el agente o el ejecutante-experimentador sabe y está consciente de cómo es el entendimiento de una lengua, el cual se diferencia notoriamente de la operación algorítmica de manipular símbolos" (González 2012:3). Entonces, la operación descrita por la Habitación China requiere de un experimentador que coteje si efectivamente hay entendimiento lingüístico.

El internalismo de Searle pone acento en quién se sitúa "dentro de la habitación" y, en particular, en el mencionado entendimiento lingüístico. De este modo, el *juicio* que se hace de la evidencia en la Habitación China, es de carácter interno. Es decir, hay un cambio de foco en la atención, desde lo estipulado por Turing y su Juego de la Imitación: desde los interrogadores en este, a Searle dentro de la Habitación. Es, por tanto, posible concluir que el experimento mental de Searle tiene una inspiración internalista cartesiana, a pesar de que el mismo le achaca a Descartes ser "un desastre filosófico" (Searle 2004:13). Así, e independiente de esta falacia *ad hominem*, es claro que el internalismo de Searle tiene mucho de cartesianismo, el cual nos lleva a cuestionar si efectivamente una máquina tiene estados mentales y es inteligente.

La competencia Loebner, en que desde 2000 se lleva a cabo el Juego de la Imitación tal como lo describió Turing, es tan polémica como su Test, puesto que no permite discernir con *certidumbre* la presencia de estados mentales en máquinas. El ejemplo de Descartes, por lo mismo, anticipa un problema, una discusión en la Inteligencia Artificial, a saber, cómo la mente es detectable para interrogadores que cuentan con evidencia puramente observacional. Si el presente análisis del Test de Turing es correcto, ello significaría un problema para realizar un programa de investigación basado en él. En efecto, y tal como algunos diagnostican, el Test parece llevar a un callejón sin salida respecto del reemplazo *definitivo* de la pregunta "¿Pueden pensar las máquinas?"

En las dos últimas secciones he agregado otra dificultad a dicho reemplazo, base del proyecto de investigación de Turing en Inteligencia Artificial, a saber, el criterio internalista de los interrogadores para decidir si se está frente a una máquina o a un ser humano hace que no resulte claro si el reemplazo definitivo de la pregunta resulta, de hecho, factible. Más aún, dicho criterio parece consistente con una práctica típicamente humana: dudar de la existencia de estados mentales en todo aquello que no es humano, como las máquinas. Esto, tal como he argumentado aquí, agrega un nuevo elemento a la discusión sobre el famoso y controvertido Test de Turing, uno inspirado por Descartes, a propósito de su caso de los hombres con capas y sombreros que ve por la ventana.

Conclusión

Las raíces del problema de las otras mentes pueden rastrearse hasta un ejemplo planteado por Descartes en las *Meditaciones Metafísicas*: los hombres con capa y sombrero que "ve" a través de la ventana. Dicho ejemplo, analizado a propósito del carácter inobservable de la mente y de que las palabras son engañosas en este respecto, muestra que es solo la razón, no los sentidos, lo que permite establecer el ser de las cosas, o como lo pone Descartes, las ideas claras y distintas de las mismas. En el caso de la cera, esta es una inspección del espíritu, mientras que en el de los hombres, estos son *res cogitans* de las cuales hay ideas claras y distintas también. Solo dichas ideas garantizan acceder a conocimiento *cierto*, es decir, el de aquél fundamentado en la razón, la cual duda acerca de lo captado por los sentidos.

En este artículo he intentado mostrar que Descartes propone, en el caso de los hombres que ve por la ventana, un criterio internalista para descubrir otras mentes. En efecto, el francés *juzga* que está frente a hombres, porque si fuera por los sentidos solamente, los sombreros y capas podrían ocultar meros autómatas. Desde el punto de vista de lo que se ve *solamente*, no hay diferencia entre los seres humanos y los autómatas. Descartes considera que estos son cuerpos que se mueven, pero incapaces de lenguaje y acción inteligente, los dos signos de pensamiento e inteligencia. Dichos signos, pese a que se observan, se juzgan internamente como característicamente humanos. Así, no es mediante los sentidos que se establece que se está en presencia de seres humanos, sino mediante la razón.

A diferencia de la mayoría de los comentaristas, en este ensayo he enfatizado de qué forma los interrogadores en el Juego de la Imitación son cruciales para juzgar el destino de tal propuesta. En efecto, el matemático Alan Turing no parece haber reconocido que los interrogadores de su Juego de la Imitación también aplican un criterio internalista para la existencia de estados mentales. De hecho, las máquinas programadas son *descubiertas*, o bien los seres humanos son *identificados*, con base en lo que los seres humanos típicamente responderían a las preguntas formuladas por los interrogadores. Es decir, estos *saben* y están conscientes de qué se debería responder frente a las preguntas formuladas.

De esta forma, en las propuestas de Descartes y Turing hay criterios internalistas para juzgar la existencia de otras mentes. Ello acontece incluso si hay factores externos, observables, que se podrían considerar en la decisión tomada. En el caso de Descartes, son las cuestiones relacionadas con el lenguaje y la acción inteligente humana; en cambio, en Turing es la conducta lingüística de los interrogados. Pero, tal como he argumentado aquí, es el estar consciente de dichos factores lo que permite *decidir* si se está frente a mentes o máquinas. Ello, si bien es un elemento esperable en la filosofía cartesiana, representa una sorpresa y una complicación extra en la propuesta de Turing, especialmente si esta tiene como objetivo reemplazar *definitivamente* la pregunta "¿Pueden pensar las máquinas?".

El Test de Turing, uno de los métodos en Inteligencia Artificial más controvertidos y más discutidos, parece no lograr el reemplazo de la filosófica pregunta: "¿Pueden pensar las máquinas?". Debido a esta dificultad, han surgido distintos intentos de mejora de la propuesta original, como el Test Total de Turing o el Test Totalmente Verdadero de Turing. Si bien discutir tales propuestas va más allá de los límites de este ensayo, puedo aseverar que ellas solo corroboran que el Juego de la Imitación toca un viejo problema filosófico, uno que consiste en el conocimiento cierto sobre la detección de estados mentales ajenos. Un problema que revive ocasionalmente, y que representa una dificultad para el programa de investigación de Turing, basado en la imitación de capacidades consideradas, por Descartes al menos, exclusivamente humanas: el lenguaje, el pensamiento y, sobre todo, la inteligencia.

Bibliografía

Crane, T. 2003. *The mechanical mind: A philosophical introduction to minds, machines and mental representation*. Abingdon: Routledge.

Descartes, R. 1994. *Discurso del método*. Madrid: Alianza.

Descartes, R. 1985. *Principles of philosophy.* En: J. Cottingham, R. Stoothoff y D. Murdoch *The philosophical writings of Descartes,* Vol. I., pp. 160-212. New York: Cambridge University Press.

Descartes, R. 1977. *Meditaciones metafísicas*. Madrid: Alfaguara.

González, R. 2015. ¿Importa la determinación del sexo de los participantes en el test de Turing? *Revista de Filosofía Aurora* 27(40): 277-295. doi: 10.7213/aurora.27.040.AO02

González, R. 2012. La pieza china: ¿un experimento mental con sesgo cartesiano? *Revista Chilena de Neuropsicología* 7(1): 1-6. doi: 10.5839/rcnp.2012.0701.02

Hyslop, A. 2016. Other minds. En: E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*. Stanford: Stanford University. http://plato.stanford.edu/archives/spr2016/entries/other-minds/

Searle, J. 1990. Mind, brains and programs. En: M. Boden (ed.) *The philosophy of artificial intelligence*, pp. 67-88. Oxford: Oxford University Press.

Searle, J. 2004. *Mind: A brief introduction*. Oxford: Oxford University Press.

Turing, A. 1951. Can digital computers think? Tipeo de una entrevista radial publicada en el Tercer Programa de BBC, 15 de mayo de 1951. Referencia de los archivos Turing número B.5. En: S. Shieber (ed.) *The Turing test: verbal behavior as the hallmark of intelligence*, pp. 111-116. Cambridge: MIT Press.

Turing, A. 1950. Computing machinery and intelligence. En: M. Boden (ed.) *The philosophy of artificial intelligence,* pp. 40-66. Oxford: Oxford University Press.

Recibido el 5 May 2016
Aceptado el 1 Jun 2016