Máquinas sin engranajes y cuerpos sin mentes
¿Cuán dualista es el Funcionalismo de Máquina de Turing?[1]

Rodrigo González
rodgonfer@gmail.com

Departamento de Filosofía y Centro de Estudios Cognitivos
Facultad de Filosofía y Humanidades
Universidad de Chile,
Av. Ignacio Carrera Pinto 1025, Ñuñoa, Stgo., Chile.

Resumen
En este trabajo examino cómo el Funcionalismo de Máquina de Turing resulta compatible con una forma de Dualismo, lo que aleja a la IA clásica o fuerte del Materialismo que la inspiró originalmente en el siglo XIX. Para sostener esta tesis, argumento que efectivamente existe una notable cercanía entre el pensamiento cartesiano y dicho Funcionalismo, ya que el primero afirma que es concebible/posible separar mente y cuerpo, mientras que el segundo sostiene que no es estrictamente necesario que los estados mentales se realicen en las propiedades físicas de engranajes y máquinas reales.

Abstract
This article deals with how Turing Machine Functionalism turns out to be compatible with a form of Dualism, which involves that strong AI is not close to the original Materialism that inspired it in the nineteenth century. To support this thesis, I argue that there is a compelling coincidence between Descartes' philosophy and this version of Functionalism, since the former holds that it is conceivable/possible to separate mind and body, while the latter holds that it is not strictly necessary that mental states are realized by the physical properties of real cogs and machines.

Palabras clave: problema mente-cuerpo, Dualismo, Funcionalismo de Máquina de Turing, Materialismo.

Keywords*: *Mind-Body problem, Dualism, Turing Machine Functionalism, Materialism.

Introducción

La historia de la Inteligencia Artificial se remonta al siglo XIX cuando Charles Babbage y su amigo John Herschel se abocaron seriamente a resolver el problema de las tablas de cálculo. Para tal efecto, se empeñaron en crear una máquina capaz de exhibir inteligencia matemática sin la típica falibilidad humana. Algunos siglos antes, Descartes negó que cualquier máquina basada en mecanismos pudiera pensar, ser racional e inteligente, ya que

[1] Agradezco el financiamiento obtenido del proyecto FONDECYT 11080020 “Lo imaginario y lo posible: un estudio desde la experiencia concebible”.

un mecanismo no puede pensar por muy sofisticado que sea. Para el filósofo francés, los mecanismos se basan en un funcionamiento regular, predecible y finito, cuestión que hace que sean incapaces de alcanzar la variabilidad y flexibilidad de la razón, dos características que se manifiestan de manera evidente en el uso del lenguaje.

Varios de los pensadores que contribuyeron al nacimiento de la IA clásica o fuerte[2,] alrededor de 1950, simpatizaron con el revolucionario proyecto de Babbage, quien defendía el monismo materialista. Estas ideas lo llevaron a sostener que Dios era un agente material, cuestión que a su vez le significó la reprobación de un examen público en la Universidad de Cambridge (Swade 2000, p. 19). Puesto su pensamiento filosófico de una manera breve y simple, Babbage pensaba que la inteligencia estaba *en la máquina*, lo que contrasta con lo argumentado por Descartes, quien aseveraba que no era posible en principio dotar a una máquina de pensamiento y racionalidad.

Entre los pensadores que contribuyeron al nacimiento de la IA en el siglo XX destaca Alan Turing, quien definió qué era computar (Turing 1936), e ideó un test para reemplazar la pregunta de si una máquina podía en efecto pensar, el famoso y controvertido Test de Turing (Turing 1950). Este justamente se opone radicalmente a las consideraciones cartesianas respecto del pensamiento, el lenguaje y los mecanismos, en tanto supone que una máquina programada sí puede hablar y pasar dicho test.

La singularidad de este método y su incapacidad para reemplazar la pregunta sobre si las máquinas piensan han sido objeto de diversas críticas (por ejemplo, González 2007). Sin embargo, a la luz del Dualismo Cartesiano y del Materialismo clásico decimonónico resulta pertinente mencionar la diferencia que existe entre Babbage y Turing con respecto a la inteligencia. Mientras que el primero creía que la inteligencia estaba en la máquina (podría calificar como inteligencia *in re*, según la jerga escolástica), y por tanto, que no podía separarse de los mecanismos concretos, el segundo pensaba que estos eran totalmente irrelevantes para crear máquinas inteligentes, pues la inteligencia puede considerarse como una función (y sería, entonces, *ante o post rem*, según la misma terminología).

---

[2] En delante me aboco al examen del Funcionalismo de Máquina de Turing y su relación con la IA clásica o fuerte, según la terminología de Searle (1980). Pese a la importancia de otras teorías como el Conexionismo (Rumelhart *et al.* 1986) o la de Clark (2008), estas no tienen directa relación con el pensamiento cartesiano de separar mediante intuiciones modales mente y cuerpo. Tal como se adelanta en esta introducción, el interés de este trabajo consiste en examinar cuán cierta es la acusación de Searle de que la IA fuerte es una forma de dualismo encubierto a la luz de comparar esta con la famosa separación real de lo extenso y lo mental que defiende Descartes.

Contrariamente a lo defendido por Babbage, Turing propone que cualquier máquina programada o digital capaz de imitar los *inputs* y *outputs* del cerebro es, como cuestión de hecho, inteligente, y lo es con independencia de los materiales que implementen la máquina o sus mecanismos. En particular, Turing afirma que "si alguna máquina particular puede ser descrita como un cerebro, solo debemos programar nuestra máquina digital para imitar a esta y también será un cerebro" (Turing 1951, p. 112). Es decir, se postula que la inteligencia es un subproducto de la función computable de una máquina con independencia del material de esta, y que la clave de crear inteligencia consiste en diseñar máquinas digitales que imiten lo que hace un cerebro.

Tal como argumento en este trabajo, las ideas de Turing, que son fundamentales para entender la teoría funcionalista de los estados mentales (Putnam 1967) y el principio de realizabilidad múltiple, curiosamente semejan la intuición modal cartesiana según la cual es ciertamente *concebible* que existan estados mentales sin corporalidad alguna, tesis que le ha valido a la IA clásica haber sido acusada por Searle de adscribir al Dualismo. Tómese a modo de ejemplo este famoso pasaje donde Searle acusa a la IA fuerte:

> […] Ciertamente, la IA solo tiene sentido dado el supuesto dualista de que, en lo que concierne a la mente, el cerebro no importa. En la IA (y en el Funcionalismo también) lo que importa son los programas, y estos son independientes de su realización en máquinas; sin duda, en lo que respecta a la IA, el mismo programa podría realizarse por una máquina electrónica, una substancia mental cartesiana, o el espíritu hegeliano del mundo. La característica más sorprendente que he descubierto examinando estos problemas es que muchos investigadores de la IA resultan choqueados por mi idea de que los fenómenos mentales humanos podrían ser independientes de las propiedades físico-químicas de los cerebros humanos reales. Pero si Ud. piensa la idea por un momento, puede apreciar que no debería haberme sorprendido, ya que a menos que se acepte una forma de Dualismo, la IA no tiene ninguna posibilidad. El proyecto consiste en reproducir y explicar lo mental diseñando programas, pero a menos que la mente no solo sea conceptual sino empíricamente independiente del cerebro, Ud. no puede llevar a cabo dicho

> proyecto, pues *el programa es completamente independiente de cualquier realización* (Searle 1980, p. 86, énfasis mío).

A la luz de las críticas de Searle, en este trabajo me propongo mostrar que esta acusación, reiterada en varias otras ocasiones (Searle 1990, 1992, y 2002), no es infundada, pero requiere de mucha más precisión en cuanto a lo que implican el Dualismo cartesiano y el Funcionalismo de Máquina de Turing. Mientras que la IA se basa en el carácter matemático y algorítmico de esta máquina, la postulación de Descartes de la existencia de estados mentales sin corporalidad se funda en que se concibe *clara y distintamente* la posibilidad de que la mente pueda existir sin lo corporal. Existe, por lo tanto, una diferencia entre lo que implica la realizabilidad múltiple, esto es, que la implementación material se lleve a cabo en cualquier medio y la posibilidad lógica explorada por Descartes mediante la separación *real* de substancias, esto es, mediante una intuición modal, que distingue clara y distintamente entre lo mental y el pensamiento, por una parte, y lo corporal y el automatismo de máquinas y mecanismos, por otra.

En la primera sección, caracterizaré brevemente la visión cartesiana que niega que una máquina pueda tener la inteligencia y racionalidad de un humano. En la segunda, examinaré de qué manera los defensores de la IA clásica conciben que una máquina programada si podría tener mente, con independencia de sus materiales. El nexo entre la intuición modal cartesiana de que puede existir mente sin cuerpo y la irrelevancia de los materiales de una máquina programada será examinado en la tercera sección de este trabajo.

1. El lenguaje como signo de racionalidad y pensamiento

Para establecer que una máquina es incapaz de pensamiento, Descartes se centra en el uso de signos lingüísticos. Según el filósofo francés, los animales son incapaces de usar dichos signos, pues toda su conducta y movimientos corporales se explican por ciertas estructuras mecánicas internas. Existe una notoria diferencia entre los signos naturales, entre los que figuran las respuestas naturales frente a un estímulo determinado, y los signos lingüísticos, que son convencionales e involucran conciencia. Esto es, mientras que existen *signos*

naturales como respuestas automáticas y genéricas al ambiente en una especie determinada, y que por ejemplo hacen que un animal gima de dolor al ser golpeado, los *signos convencionales* lingüísticos nunca son empleados mecánica o automáticamente.

La diferencia entre el uso de signos por parte de animales y humanos no es de grado, sino de naturaleza, tal como Descartes enfatiza en la siguiente carta a Mersenne, en 1629:

> Con respecto a las palabras que significan de manera natural, acepto como una explicación válida que cualquier cosa que impacte nuestros sentidos nos lleva a emitir algún sonido; por ejemplo, si somos golpeados, nos hará quejarnos; si alguien nos hace algo que cause placer, nos hará reír; y que los sonidos que uno emite, al quejarnos o reír, son similares en todas las lenguas. Pero cuando veo el cielo o la tierra, no resulto forzado a llamarlas de una manera u otra y creo que esto sería el caso incluso si estuviera en un estado de inocencia adánica (Descartes 1629, en Clarke 2003, p. 163)

No obstante, es preciso ser cuidadoso con este pasaje, pues no significa que los animales no usen "sus propias palabras", entendiendo estas como los signos naturales a los que resultan forzados por estimulaciones externas. En realidad, el uso del lenguaje por parte de los humanos muestra de manera *inequívoca* la existencia de pensamiento, capacidad que va de la mano de la flexibilidad de la razón y de la voluntad libre para elegir palabras y comunicar ideas. Cuando alguien golpea a un animal y emite un sonido de queja de manera automática e involuntaria, no existe tal libertad, sino que los famosos "espíritus animales" mueven a ciertas disposiciones naturales de los órganos del cuerpo, y existe conducta sin la participación de la voluntad; luego, la conducta animal no está mediada por ninguna decisión racional.

A su vez, los espíritus animales explican las emociones espontáneas, en las cuales no se ejerce ninguna decisión. Para esclarecer a qué se refiere con los espíritus animales, Descartes asevera que "son como el sutilísimo viento, o más bien como una vivísima y purísima llama, la cual asciende de continuo muy abundante desde el corazón al cerebro y se corre por los nervios a los músculos y pone en movimiento todos los miembros"

(Descartes 1996a, p. 54). Por este motivo, resulta natural la elucidación de los espíritus animales en términos de las leyes de la mecánica, y la natural oposición entre estas y el pensamiento.

La emisión de palabras, por otra parte, no implica una referencia natural a lo significado, que es producto de una convención, punto que se enfatiza en *Tratado del Mundo, o Sobre la Luz* de la siguiente forma:

> Ahora, si las palabras ―las cuales tienen significado solo como resultado de convención humana― son suficientes para hacernos pensar acerca de las cosas que nos recuerdan estas, ¿por qué no es posible que la naturaleza pudiera haber establecido un signo particular que nos hiciera tener la sensación de luz, incluso aunque dicho signo no contuviese nada que recordase a la sensación? ¿Y no es esta la manera en la que se ha establecido que la risa o las lágrimas, nos hacen leer el gozo o la pena en la cara de las personas? (Descartes 1996b, p. 4)

Tales preguntas tienen una respuesta clara: el nexo entre signo natural y conducta automática no está mediada por el entendimiento y la razón. No obstante, y pese a lo sostenido respecto de los espíritus animales y los signos convencionales lingüísticos, Descartes evaluó posibilidad de que una máquina pudiera hablar en el *Discurso del Método* de la siguiente forma:

> Uno podría concebir una máquina que fuese hecha de tal forma que respondiera palabras, y que incluso dijera palabras como respuesta a las acciones físicas que causan cambios en sus órganos ―por ejemplo, si alguien la tocase en un lugar específico, podría preguntar qué quiere uno que diga, o si fuese tocada en otro lugar, podría quejarse diciendo que se le está dañando, y así sucesivamente (Descartes 1996a, p. 56)

Lo que hace que una máquina no pueda tener el pensamiento humano no es la incapacidad de reaccionar, ya que los animales lo hacen, y cuentan como máquinas complejas. Por el contrario, existe una habilidad específica humana que muestra de manera inequívoca la

existencia de pensamiento, a saber, la capacidad para articular palabras de manera no mecánica. Justamente, la inhabilidad para articular lenguaje de una manera no programada o mecanizada y, por tanto, que puedan ser objeto de predicciones, indica que animales y máquinas no son pensantes. Descartes destaca este punto agregando inmediatamente que "[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 […] (Descartes 1996a, pp. 56-57)". Y agrega que dado que las máquinas no actúan de acuerdo con su entendimiento, sino por la entera disposición de sus órganos, podrían hacer algunas cosas que hacen los humanos bien, y otras no.

Justamente, la naturaleza *causal y mecánica* de las reacciones de animales y máquinas cuentan para Descartes como *el primer signo que muestra por qué éstas son incapaces de pensamiento*. Y este signo se explica porque el uso del lenguaje *no* puede explicarse mecánicamente, y tampoco ostenta un vínculo causal entre estímulos ambientales y respuestas lingüísticas, con reordenación de palabras para significar lo mismo de variadas maneras.

No obstante, estas disquisiciones Cartesianas en que compara a los humanos con animales y máquinas, y que enfatizan la importancia del lenguaje, van más allá todavía, pues el uso de palabras es considerado un signo inequívoco de racionalidad, lo que contaría como *el segundo signo que indica la existencia de pensamiento.*

Ahora bien, la racionalidad no solo es definida por Descartes en función de la habilidad para aprender a usar palabras y signos convencionales, sino además por la capacidad de actuar inteligentemente en función de conocimiento, idea que está ligada a que mientras "la razón es un instrumento universal que puede usarse en toda clase de situaciones, los órganos necesitan cierta disposición particular para cada acción específica" (Descartes 1996a, p. 57).

En vista de que el número de respuestas lingüísticas apropiadas es indefinidamente grande, y de que el número de estímulos lingüísticos al que pueden responder los humanos es igualmente grande, se sigue que es imposible para una máquina y para un animal almacenar y reproducir respuestas inteligentes que imiten *la flexibilidad y variabilidad de la razón humana*, pues ambos poseen una capacidad de programación limitada. Solo el

entendimiento es capaz de determinar qué acción es la más adecuada dentro de un contexto específico y, por tanto, actuar de manera inteligente.

Sin embargo, se debe enfatizar que estas consideraciones mecanicistas cartesianas sobre los animales no significan que estos no sientan, ni que carezcan de conciencia, sino que, como certeramente aclara Cottingham (1998, pp. 225-226), solo implican que los animales son autómata que no piensan como nosotros, que carecen de lenguaje, y que no tienen autoconciencia.

Así, todo ser que sustente su actuar en mecanismos fijos y determinados, será incapaz de usar lenguaje, pensar y ser racional, lo que contrasta notoriamente con el proyecto decimonónico de la IA, concebido inicialmente desde una perspectiva monista materialista por Babbage, e implementado por Turing a partir de las nociones de algoritmo y computador digital en el siglo XX.

2. La Máquina de Turing y la inteligencia: mecanismos sin engranajes concretos

Entre los fundadores de la Inteligencia Artificial durante la primera mitad del siglo XX, sin duda destaca Alan Turing, quien contribuyó a cuestionar el pensamiento cartesiano sobre por qué no es plausible atribuir inteligencia a una máquina programada. Su contribución se basa en dos supuestos cruciales: i) es posible mecanizar el pensamiento; ii) es posible crear inteligencia mediante procedimientos efectivos o algorítmicos.

Estas tesis de Turing se remontan al *Entscheidungsproblem*, tratado por David Hilbert en la década de 1930. Dicho problema trata sobre la posibilidad de establecer un procedimiento algorítmico para decidir si los enunciados matemáticos son demostrables, o si una fórmula lógica de primer orden es universalmente válida. La investigación de Turing (1936), además de demostrar que no es posible encontrar tal procedimiento, permitió comprender la noción de computar y también cómo una máquina podría realizar procedimientos equivalentes al pensamiento, una prerrogativa exclusiva de los humanos según Descartes. Pese a esto, Descartes y Turing coinciden en el carácter multipropósito y flexible de la inteligencia. Turing, por ejemplo, concibió que las máquinas programadas sí son capaces de resolver cualquier problema e incluso aprender por medio de la

computación de sus representaciones simbólicas. Estas computaciones tienen lugar mediante las operaciones típicas de una Máquina de Turing.

Esta es la idealización matemática de un dispositivo mecánico que opera sobre secuencias de unos y ceros y mediante estados determinados o discretos (Turing 1936, pp. 231-5), cuestión que se liga al sentido intuitivo de computar (Turing 1936, 250-3). Tal acción involucra una cantidad finita de estados discretos, o la implementación de un algoritmo, cuyos pasos, también finitos, son calculables mediante lápiz y papel y sin la participación de conciencia. El término del procedimiento descrito por el algoritmo se alcanza cuando se encuentra la solución a un problema. Justamente, un "computador humano", que emplea lápiz y papel, puede definirse en términos de una Máquina de Turing.

Los algoritmos, fundamentales para entender qué es una Máquina de Turing, eran ya conocidos en la antigüedad. Un esclarecedor ejemplo es el procedimiento planteado por Euclides para encontrar el máximo común denominador de dos números al dividir sucesivamente la segunda cifra divisora por el remanente de la división entre la primera y la segunda cifra, hasta que éste sea 0 (para detalles de este ejemplo ver Penrose 1999, pp. 40-45). Tal procedimiento sistemático es *efectivo*, pues describe una serie finita de pasos y se detiene cuando se alcanza un resultado específico.

Pero, ¿cuál es la relación entre un algoritmo y una Máquina de Turing? Esta máquina fue concebida por Turing como una manera de formalizar términos como "procedimiento mecánico" y "máquina", puesto que dicho dispositivo simplemente ejecuta un cálculo definible de manera finita y recursiva, esto es, implementando o llevando a cabo un algoritmo. Por lo mismo, es una máquina idealizada, con una cantidad de estados discretos posibles, los cuales son finitos aunque pueden ser muchísimos. Pese a esta aparente restricción, no posee límite con respecto a los cálculos posibles, ya que tiene un conjunto de instrucciones que se ejecutan con independencia del tamaño de los números a calcular. Las entradas o *inputs* son secuencias de símbolos de un alfabeto finito, y se utiliza una capacidad de almacenaje de información externa, la cual es usualmente descrita como papel para realizar los cálculos y producir el *output.*

Asimismo, no se supone que la máquina internalice los datos externos o los cálculos, sino que trata con cálculos u operaciones inmediatas. El tamaño ilimitado del *input* versus la finitud de los estados o pasos para los cálculos corrobora (aunque no

implica) que las Máquinas de Turing son *idealizaciones matemáticas*. Penrose insiste sobre este punto de la siguiente manera:

> Es la naturaleza ilimitada del input, del espacio de cálculo, y del output lo que nos indica que estamos considerando *una idealización matemática* en vez de algo que puede construirse en la práctica […] Las maravillas de la tecnología de los computadores modernos nos han provisto de dispositivos electrónicos de almacenamiento los que, ciertamente, pueden tratarse como ilimitados para la mayoría de los propósitos (Penrose 1999, p. 47, énfasis mío)

Usualmente, la Máquina de Turing se representa mediante una cinta infinita dividida en celdas, con una cabeza lecto-escritora, que se mueve hacia derecha e izquierda a través de la cinta, y que "recuerda" alguno de los símbolos leídos o estados discretos (por contraposición a continuos). En un tiempo *t*, la cabeza que se encuentra en un estado interno (q_0, … , q_n), lee el símbolo de la celda de la cinta (b_1, … , b_n). En función del estado interno, de lo leído, y del micro código o programa de la cabeza, esta mantiene el símbolo o lo borra y escribe otro. Luego, se detiene o se mueve a otra celda, continuando las computaciones hasta que se detenga la cabeza, lo cual marca el final del procedimiento.

Así, la Máquina de Turing tiene 0 y 1 como símbolos en las celdas, y su conducta está totalmente determinada por la tabla de la máquina o el programa. Por ejemplo, la máquina, al estar en un estado interno y leer el símbolo de una celda, tiene un input I_n, el cual consiste en el par <estado interno, símbolo leído>. Luego, en virtud de I_n y del programa trata de determinar una tripleta de *output*, esto es, <escribe símbolo, movimiento (o detención) y nuevo estado interno>. Si la máquina no es capaz de determinar la tripleta, se detiene, y también lo hace cuando la computación ha llegado a su fin. Sobre la base de O_n, la máquina escribe un símbolo (1 o 0) en la celda, se mueve a la izquierda o derecha (o se detiene ahí) e ingresa a un nuevo estado. Eventualmente, la máquina vuelve al primer paso, si no ha llegado a detenerse luego de computar la solución del problema.

Toda la conducta de la máquina se estipula mediante enunciados condicionales, y del programa con la lista finita de reglas ordenadas en pares de <entrada, salida>. Las reglas son expresadas mediante fórmulas "Si …, entonces…", las que sirven para

conformar quíntuplos de operaciones de la forma <estado actual, leer símbolo, escribir símbolo, moverse (o detenerse) y nuevo estado>. Una cuestión importante de tener en consideración es que los estados de la máquina, las colecciones de símbolos, y las acciones, como borrar, escribir y moverse, son además de *finitas*, *discretas* y *distinguibles*. En consecuencia, no es posible que una máquina esté en dos estados a la vez, o que cuente con una cantidad infinita de estados posibles.

Los procedimientos de una Máquina de Turing son algorítmicos y efectivos. ¿Pero qué quiere decir efectivo en este contexto? Según Copeland (2002), un procedimiento *M* es efectivo si y solo si: i) *M* se concibe mediante una lista finita de instrucciones, las que son expresadas por una lista finita de símbolos; ii) *M* se ejecuta sin errores en una serie de pasos finitos; iii) *M* puede ejecutarse por parte de un humano que emplee lápiz y papel para los cálculos; iv) La ejecución de *M* no requiere de conciencia para realizar los cálculos. A su vez, si un procedimiento es efectivo y, por tanto, su función es computable, lo es por una Máquina de Turing.

El poder de una Máquina de Turing radica en que puede, en principio, simular cualquier sistema algorítmicamente calculable. Cualquier máquina que implemente un procedimiento efectivo *M* podrá ser imitada por una Máquina de Turing, y esto es supuestamente crucial para la construcción de máquinas programadas capaces de exhibir la misma inteligencia de un humano. En efecto, el proyecto de la IA clásica o fuerte contempla que máquinas programadas puedan ejecutar tareas y resolver problemas que implican la exhibición de actividad mental inteligente mediante procesos algorítimicos.

Tradicionalmente, se ha interpretado que la independencia de los materiales de la Máquina de Turing cuenta como un importante antecedente del *principio de realizabilidad múltiple*. Puesto de manera simple, este principio señala que “ser una máquina X” es funcionar como esa máquina y, así, los materiales son irrelevantes, pues qué hace y cómo funciona la máquina es lo *único* que importa.

Que “cuente a favor de” no significa que la Máquina de Turing implique el principio de la realizabilidad múltiple, ya que ambas tesis son compatibles con que exista una sola realización material de la inteligencia, por ejemplo, el cerebro, y que este sea una Máquina de Turing. Incluso, resulta al menos lógicamente posible que no exista ninguna implementación del programa ejecutado por dicha Máquina. En este sentido, la IA clásica

asume que es posible concebir un procedimiento algorítmico cuya sistematicidad, flexibilidad y complejidad sea equiparable a la inteligencia humana, pese a que no es necesario replicar las propiedades físicas del cerebro. Puesto de una manera simple, con lo anterior se asume que la inteligencia es la función o programa del cerebro, el cual puede en principio imitarse por una Máquina de Turing.

La irrelevancia del *hardware* también se explica porque, tal como Cleland (1993) manifiesta en su crítica al Funcionalismo, los nexos entre los pasos finitos de un programa, o de un procedimiento efectivo *M* que ciertamente puede ejecutar cualquier Máquina de Turing, no tienen nada que ver con la causalidad. Es decir, un procedimiento efectivo vincula *formalmente* todo los pasos. Según ella, la crucial y reveladora diferencia entre cerebros y programas queda de manifiesto entonces al comparar estos con una receta de cocina, que cuenta como un procedimiento efectivo mundano. Mientras que en un programa el vínculo entre los estados de la máquina opera por sintaxis, esto es, los pasos están conectados mediante reglas y necesariamente, la receta requiere de eventos materiales y de causalidad para producir el resultado final, y sus pasos son, por lo mismo, *contingentes* y dependientes de las leyes físicas del mundo en que se lleven a cabo los mismos.

De hecho, es interesante la conexión que existe aquí entre la variabilidad de los pasos de una receta (por ejemplo, para hacer un queque) y la concepción de la IA clásica como flexible y mecánico-abstracta. En efecto, si una Máquina de Turing puede imitar los estados posibles del cerebro, lo hará de acuerdo con la tabla de máquina, y sus secuencias sintácticas debieran establecerse *con necesidad y de modo unívoco*, que es precisamente como dichas máquinas implementan procedimientos efectivos. Esta posible complicación es prevista por Putnam, quien mantiene que los estados del cerebro no deben caracterizarse físicamente, sino por las transiciones de una Máquina de Turing que se comporte como un autómata probabilístico (Putnam 1967, p. 162, y 1973, p. 75), i.e., con transiciones entre estados regulados mediante probabilidades y no de forma determinista.

Más importante todavía es que los defensores de la denominada IA fuerte aseveran que la mente puede modelarse computacionalmente, ya que una máquina que opere como un cerebro *es* un cerebro artificial. Es decir, la mente humana no sería más que un computador digital apropiadamente programado, tesis que resulta complementaria con lo postulado por la Ciencia Cognitiva de que el cerebro es una máquina sintáctica operando

una máquina semántica (Block 1990, p. 267). Es decir, los seguidores de la IA y de la Ciencia Cognitiva sostienen que existen ciertas estructuras simbólicas en el cerebro que se correlacionan y explican las actividades del pensamiento y de la inteligencia humana.

De esta forma, la IA fuerte asume que cualquier problema que requiera inteligencia puede resolverse con la aplicación del algoritmo adecuado, cuestión que es complementaria con la hipótesis de la Ciencia Cognitiva ya descrita. Si a ambas ideas se suma la irrelevancia del material en que se realiza un programa, una idea peculiar del Funcionalismo de Máquina de Turing, la separación cartesiana de mente y cuerpo parece más cercana. Tal como Searle argumenta en diversas ocasiones (1980, 1990, 1992 y 2002), existe cercanía entre la IA fuerte y el Dualismo cartesiano. El siguiente pasaje es bastante ilustrativo:

> La IA es una extraña mezcla de Conductismo y Dualismo. Es conductista en su aceptación del Test de Turing, pero en un nivel más profundamente filosófico es dualista, porque rechaza la idea de que la conciencia y la intencionalidad son fenómenos biológicos ordinarios como la digestión. En palabras de Dennett y Hofstadter, debemos pensar en la mente como, "un tipo de cosa abstracta cuya identidad es independiente de una encarnación física" (Dennett y Hofstadter 1981, p. 15, citado en Searle 2002, p. 57).

Es curioso que nunca se haya analizado de dónde provenía tal supuesto Dualismo de la IA fuerte o clásica. Efectivamente, en ésta existe un importante vínculo entre la irrelevancia del material de una máquina digital y la idea de la inmortalidad del alma, cuestión que nuevamente la acerca a Descartes y al pensamiento de que es concebible y posible separar mente y cuerpo. Pero, aunque el Funcionalismo de Máquina de Turing no *implica* una forma de Dualismo, el principio de realizabilidad múltiple sí resulta compatible con una de sus formas.

Pese a esta interesante cercanía entre la IA clásica o fuerte y Descartes, la idea central que distingue a ambos es que la inteligencia y el pensamiento no pueden ser mecanismos, incluso si estos tienen la naturaleza formal de una Máquina de Turing. El análisis de esta incompatibilidad es justamente el objetivo de la última sección.

3. ¿Máquinas sin engranajes=mentes sin cuerpo? Sobre el peculiar estatus metafísico de las máquinas de la IA fuerte

Los desarrollos de la Inteligencia Artificial tuvieron interesantes repercusiones en la Filosofía de la Mente contemporánea. El Funcionalismo nace como una teoría acerca de la naturaleza de los estados mentales alternativa a la teoría de la identidad, la cual típicamente reduce estados mentales tipo a estados neurales tipo, es decir, intenta reducir lo mental a lo físico. De acuerdo con este *fisicismo* reductivo, cualquier propiedad mental (e.g. tener dolor) debería identificarse o reducirse a una propiedad física (e.g. la activación de las fibras C), y la existencia del primer tipo de propiedades no podría darse sin la existencia del segundo tipo de propiedades, ya que existiría una relación de identidad y dependencia entre ambas. Una serie de objeciones modales (Kripke 1980, pp. 144-146) al *chauvinismo neural* que se sigue de este intento de reducción materialista, precisamente basadas en intuiciones modales cartesianas, pusieron en entredicho el Materialismo tipo-tipo, lo que ayudó a catapultar al Funcionalismo, con su ulterior apelación a los estados discretos de una Máquina de Turing en tanto capaces de caracterizar funcionalmente los estados mentales. La inspiración cartesiana del argumento de Kripke en contra de la Teoría de la Identidad puede apreciarse en el siguiente pasaje de su célebre libro *El Nombrar y la Necesidad*:

> Descartes y algunos seguidores argumentaron que una persona o mente es distinta de su cuerpo dado que la mente podía existir sin el cuerpo. Pudo haberse argumentado a favor de la misma conclusión a partir de la premisa de que el cuerpo podría haber existido sin la mente. Ahora bien, la respuesta que considero sencillamente inadmisible es la que acepta gustosamente la premisa cartesiana en tanto niega la conclusión cartesiana. Sea “Descartes” un nombre, o designador rígido de su cuerpo, de determinada persona, y sea “B” un designador rígido de su cuerpo. Entonces, si Descartes fuera efectivamente idéntico a B, la supuesta identidad al ser una identidad entre dos designadores rígidos, sería necesaria y Descartes no podría existir sin *B* y *B* no podría existir sin Descartes. El caso análogo

> no es ninguna manera comparable al supuesto caso análogo del primer director general de Correos y el inventor de los lentes bifocales (Kripke 1980, pp. 150-1).

Lo crucial del argumento kripkeano es sí es posible concebir estados mentales sin estados neurales, puesto que puede darse el caso del dolor sin la activación de las Fibras C o viceversa. Esto socavaría las bases de la Teoría de la Identidad y favorecería el *fisicismo* no reductivo de los funcionalistas. Explicado de una manera simple y no técnica, el Funcionalismo afirma que un estado mental es solo un estado con causas y efectos típicos (Lewis 1966, p. 153), y que se debe distinguir entre el rol causal de un estado mental y el ocupante, o la función que desempeña un estado mental en la economía funcional de un organismo y su realización física. Un rol podría tener varios ocupantes y realizadores. Por ejemplo, el dolor de cabeza sería un conjunto de causas y efectos conductuales, que podría realizarse en distintos sustratos materiales, no siendo las propiedades físicas fundamentales ni reductivas de las propiedades mentales.

Una variante de esta aproximación sostiene que un estado mental es compatible con una serie de predicados físicos diferentes (Putnam 1967, p. 162), tesis que da lugar al Funcionalismo de Máquina de Turing. Esta variante profundiza la idea anterior, al sostener que los estados mentales pueden reducirse a los estados de una tabla de máquina, o a la disyunción de dichos estados. Originalmente, la tesis del Funcionalismo de Máquina de Turing surgió como una aclaración de qué objetos pueden compartir propiedades psicológicas y, por tanto, a cuáles se les pueden aplicar tales predicados con independencia de sus propiedades físicas.

Esta teoría, que versa sobre la naturaleza de los estados mentales, adscribe como se ha adelantado, al principio de realizabilidad múltiple, tesis según la cual un estado mental tipo al que se le aplica un predicado psicológico, puede tener distintas realizaciones en distintas propiedades físicas. Aunque esta diferencia es importante, los organismos que tengan estas últimas compartirán el mismo perfil causal (i.e., el rol funcional) de tal estado mental.

Tal como se caracterizó en la sección anterior, la Máquina de Turing permitiría describir computacionalmente los estados mentales, realizándose en una gran variedad de materiales. Por lo mismo, no importa que sus estados no sean especificaciones de estados

neurales tipo, tal como el *chauvinismo neural* de la teoría de la identidad de tipos defiende. Por el contrario, la caracterización funcional de los estados mentales mediante autómata finito permitiría obviar el material y los realizadores de tales estados, lo que ocurre pues los autómata son definidos de manera similar a una Máquina de Turing, "excepto que las transiciones entre los 'estados' se permiten con varias probabilidades en vez de determinísticamente" (Putnam 1967, p. 162).

Tendiendo presente estos argumentos en contra del *chauvinismo neural*, la mente en su conjunto podría verse como una Máquina de Turing con estados probabilísticos. Si esta caracteriza funcionalmente ciertos estados, podrían existir una gran variedad de implementaciones de la máquina que no tuvieren nada que ver con cerebros y sistemas nerviosos, siempre y cuando se preservara la economía funcional y la caracterización de los estados mentales en términos de causas y efectos. En consecuencia, el funcionalista de Máquina de Turing propone que la organización funcional mediante perfiles causales, y no la realización material de los estados mentales, es crucial para caracterizar estos.

Sin embargo, esta afinidad entre una Máquina de Turing y la caracterización de los estados mentales mediante estados funcionales también permitiría la existencia de entidades concientes no-físicas, con perfiles causales compatibles al humano o animal. Desde un punto de vista conceptual al menos, es posible plantear que existen entidades concientes no-físicas con perfiles causales adecuados similares al humano, pero sin ninguna realización física. Por ejemplo, resulta al menos *concebible* la existencia de un estado mental como el dolor en seres ectoplásmicos[3] o en ángeles, en el sentido de que estos seres podrían tener un perfil causal similar al de humanos o animales y una implementación que no fuera física. Este problema para el Funcionalismo es abordado de una manera iluminadora en la siguiente caracterización de Heil, en la cual se enfatiza como esta teoría es usualmente considerada como incompatible con el Materialismo clásico:

> Si los estados de la mente son estados funcionales, y si los estados funcionales se "realizan" en criaturas concientes a través de estados con un perfil causal apropiado, esto deja abierta la posibilidad de que seres inmateriales puedan estar concientes,

[3] La presencia de espíritus y fantasmas se explica por una sustancia denominada ectoplasma, que cuenta como un tipo de implementación. Por supuesto es difícil establecer cuán material es esta sustancia, pues se comporta como una forma de energía que permite interacción causal en el mundo.

> pensar, sentir dolor. Ellos podrían poseer tales estados de la mente si pudiesen estar en estados con los perfiles causales adecuados. Una sustancia ectoplásmica, por ejemplo, o un ángel con la organización interna ectoplásmica o angelical podría pensar o sentir dolor (Heil 2004, p. 148).

Pero, ¿cómo ángeles y seres ectoplásmicos podrían tener una implementación no física en consideración de que interactúan con objetos físicos? Tal posibilidad aunque no parece implicar contradicción desde un punto de vista conceptual, pues la sustancia ectoplásmica supuestamente cuenta como un tipo de energía espiritual mundana, resulta al menos algo chocante desde el punto de vista de la física popular y del sentido común, que proponen un concepto de materia bastante concreto. Por ejemplo, tal como Dennett critica, en su alusión al Dualismo y a *Casper el fantasmita amistoso*, argumentar en base a este tipo de seres lleva a una paradoja ante la cual incluso los niños se asombran: "¿Cómo es posible que Casper traspase murallas y sea capaz de agarrar una toalla que se cae?" (Dennett 1991, p. 35). En consecuencia, si bien el Funcionalismo y el principio de realizabilidad múltiple no implican de manera obvia al Dualismo, la compatibilidad de ambos con esta concepción y con la idea de que el material no importa genera problemas metafísicos, toda vez que el Funcionalismo y otras teorías en principio *fisicistas* son justamente planteadas como alternativas materialistas a la propuesta de Descartes.

Más aún, decir que los programas son independientes de sus realizaciones físicas es una manera de sostener que el nivel sintáctico, los patrones formales, es *separable* del nivel físico, y que lo mental no está adscrito a las propiedades físicas de ningún material. La realizabilidad en múltiples materiales supone que las propiedades de estos no son relevantes para la implementación de un programa, e incluso que la realizabilidad podría concretarse a través de propiedades no físicas, tal como el ejemplo de las sustancias ectoplásmicas y ángeles sugiere. Esto, que resulta al menos conceptualmente posible sobre bases *a priori*, es crucial para apreciar cómo el Funcionalismo de Máquina de Turing se acerca al Dualismo cartesiano al proponer dos niveles de la realidad *diferentes*, a saber, el de los estados funcionales, abstractos, y el nivel de la realización de estos, que sería concreto.

Incluso, uno no requiere de seres tan polémicos como los seres ectoplásmicos y los ángeles para mostrar por qué la separación de dos niveles de la realidad involucra que uno

podría existir sin el segundo, como es el caso de los programas y sus realizaciones. Un proyecto de inversión podría ser considerado como un programa con una serie de pasos finitos regulados sintácticamente y con un objetivo central, a saber, obtener la mayor plusvalía posible. Las etapas del proyecto pueden justamente verse como los pasos de dicho programa, necesarios para la consecución del objetivo central, y por lo mismo podría considerarse un procedimiento efectivo mundano. Ahora bien, ¿es necesario que el proyecto de inversión se materialice de alguna manera, que se compren bienes tangibles y que mediante estos se obtenga la mayor plusvalía posible? En estricto rigor, no es necesario que el proyecto mismo se materialice en ningún bien tangible y concreto, ya que podría invertirse todo el capital en acciones y en otros instrumentos financieros que no tienen una realización física y que, en este sentido, serían solo mentalmente dependientes, según la terminología de Searle (1995). Incluso, el capital final obtenido mediante la ejecución del proyecto de inversión podría figurar como una cifra abstracta, sin que se materializase de ninguna forma hasta que alguien fuera a un banco e hiciera el retiro de los fondos. Por supuesto, el hecho de que un proyecto de inversión pueda contar como un programa, entendido como una serie de pasos finitos recursivos tendientes a la resolución de un problema específico, no significa que deba realizarse en computadores electrónicos específicos. Podría, por ejemplo, solo contar como una serie de ideas conectadas sintácticamente en alguna mente.

En consecuencia, existe una importante implicancia al afirmar que el nivel funcional no es reducible a un nivel material concreto, y que el primero podría realizarse de múltiples maneras en el segundo, o que incluso podría no realizarse en un objeto físico. En efecto, afirmar que las propiedades físicas no son relevantes para lograr la implementación de los estados funcionales implícitamente hace “flotar” los primeros sobre la realidad concreta del mundo físico, cuestión algo paradójica si nuevamente se tiene en consideración que el Funcionalismo pretende contar como una clase de *fisicismo* y, por tanto, como una teoría auténticamente materialista que refute a Descartes y a la idea de que lo mental no necesita estar corporeizado en un material específico, como el sistema nervioso y el cerebro.

Con respecto a la posible separación de los niveles abstractos y concretos, conviene recordar que el propio filósofo francés propone que es posible distinguir clara y distintamente que puede existir una mente sin realización física concreta, y que entonces

desde el punto de vista de lo concebible y de lo posible, luego, de las intuiciones modales, la propiedades mentales no están adscritas a las propiedades físicas, ni lo mental sería dependiente en modo alguno de lo físico. Putnam reconoce esta compatibilidad entre el Funcionalismo y el Dualismo, aunque sin tildarlo de cartesiano, en el siguiente pasaje:

> Sin embargo, los estados funcionales de los sistemas son algo distintos [a los estados físico-químicos del cerebro]. En particular, la hipótesis de los estados funcionales, ¡No es incompatible con el Dualismo! Aunque resulta patente que la hipótesis es mecanicista en cuanto a su inspiración, es un hecho marcadamente notorio que un sistema que consiste en un cuerpo y un "alma", si existe esto, podría perfectamente ser un autómata probabilístico. (Putnam 1973, p. 76)

Compárese esto con lo defendido por Descartes, quien argumenta en dos pasajes que resulta concebible separar cuerpo y mente, y que lo que prima es la existencia de lo mental, de lo que somos directamente concientes mediante ideas claras y distintas, por sobre lo corpóreo. En particular, el primer pasaje afirma que el pensamiento no está *adscrito* al cuerpo, ni a lo mecánico, pues es posible concebir la separación de lo mental y lo físico de la siguiente manera:

> Mediante el término 'pensamiento', entiendo todas las cosas de las que estamos concientes mientras nos pasan a nosotros, en cuanto tenemos conciencia de esto. Por lo tanto, el *pensar* debe identificarse no solo con entender, desear e imaginar, sino también con la conciencia de lo que sentimos. Debido a que si digo "estoy viendo, o estoy caminando, por lo tanto pienso", y tomo esto como aplicable al ver y al caminar como actividades, entonces la conclusión no es absolutamente cierta. Esto sucede porque, como a menudo pasa durante el sueño, es posible que piense que estoy viendo o moviéndome, aunque mis ojos estén cerrados o me mueva en absoluto; tales pensamientos podrían ser posibles si no tuviese cuerpo alguno (Descartes 1996c, p. 7).

En un segundo pasaje de *Las Meditaciones Metafísicas* elabora mucho más esta intuición modal, argumentando que la misma conduce a una distinción *real* entre cuerpo y mente:

> Primero, porque sé que todas las cosas que concibo clara y distintamente pueden ser producidas por Dios tal como las entiendo. Por lo tanto, del hecho de que pueda entender clara y distintamente una cosa separada de otra es suficiente para darme certeza de que las dos cosas son distintas, pues son susceptibles de ser separadas, al menos por Dios. La pregunta de qué clase de poder es necesario para causar tal separación no afecta el juicio de que las dos cosas son diferentes. Así, simplemente sabiendo que existo y que estoy viendo al mismo tiempo que absolutamente nada más pertenece a mi naturaleza excepto que soy una cosa pensante, puedo inferir correctamente que mi esencia solo consiste en el hecho de que soy una cosa pensante. Es verdad que podría tener (o que podría anticipar tener) un cuerpo que está muy cercanamente unido a mí. No obstante, tengo una idea clara y distinta de mí, en cuanto a que soy simplemente una cosa no extensa pensante; y, por otra parte, tengo una distinta idea del cuerpo en cuanto esta es una cosa extensa no pensante. Y, en consecuencia, es cierto que soy muy distinto a mi cuerpo, y que puedo existir sin este (Descartes 1996d, p. 78)

Este tipo de intuición modal cartesiana tiene, por tanto, un estrecho vínculo con cómo uno puede concebir dos cosas *separadas* y, así, que una ni está adscrita ni es dependiente ontológicamente de la otra. Desde la perspectiva del pensar, uno puede separar con claridad y distinción el cuerpo, del mismo modo en que Dios lo hace y es capaz de separar tales cosas. Es decir, en tanto es posible tener una idea clara y distinta de la existencia de la mente, *como una cosa completa* sin que exista necesariamente lo corpóreo, Descartes concluye que lo mental puede ser separado de lo corpóreo, y que existe *claridad y distinción* en la intuición que conduce a tal separación. La dirección de este argumento es, entonces, desde la certeza del pensar y de su autosuficiencia respecto de su existencia a la posibilidad de que se dé el pensar sin la existencia de corporalidad.

En vista de la imposibilidad de reducir los estados mentales a propiedades físicas, Descartes y el Funcionalismo de Máquina de Turing, base de la IA clásica, se acercan

notablemente, aunque con supuestos teóricos muy diferentes, pues el pensamiento y la racionalidad serían máximamente objetivos y, tal como se examinó en la primera sección, ninguno de estos puede ser mecánico o automático. Por otra parte, conviene también aclarar que Descartes no postuló la existencia de una multiplicidad de realizaciones posibles del pensamiento, ni menos lo hizo teniendo en consideración la máxima funcionalista de que tales realizaciones no solo son mecánicas, sino que deben ser equivalentes en función de sus inputs/outputs y estados internos.

Resulta crucial destacar esta diferencia fundamental. En efecto, mientras que Descartes afirma tener certeza de la distinción mente-cuerpo, en tanto tiene una idea clara y distinta de aquella, la IA fuerte propone que la hipótesis de que los estados mentales son estados discretos de un autómata finito no es incompatible con el Dualismo, toda vez que las similitudes en la conducta de un organismo llevan a concebir similitudes en la organización funcional, sin que sea relevante especificar e identificar detalles físicos (Putnam 1967, p. 164-5).

De esta forma, la argumentación funcionalista descansa en una aproximación externalista diferente de la cartesiana, ya que la última privilegia la introspección y la claridad y distinción de la idea que separa dos cosas de manera real. Al proponer Descartes una intuición modal de separación de mente y cuerpo, que se basa en la claridad y distinción de tal diferencia al ser escrutada por la mente, se infiere del supuesto funcionalista de la IA que es concebible que existan seres que no tengan realización material, como los seres ectoplásmicos y angelicales, los cuales *podrían* tener la misma economía funcional humana y, por lo tanto, podrían tener los mismos dolores y otros estados mentales. Esto último, en consecuencia, al menos despeja la duda de si el Funcionalismo de Máquina de Turing cuenta como una forma de Materialismo genuina, como a veces se lo caracteriza, y corrobora el distanciamiento entre la IA decimonónica de Babbage y la clásica o fuerte.

En síntesis, son justas las acusaciones de Searle a la IA clásica o fuerte de escindir el pensamiento y la inteligencia de la corporalidad y del cerebro, pues no apoya de manera consistente al monismo materialista que originalmente la inspiró en el siglo XIX con Babbage como su gran defensor.

Referencias

Block, N. (1990) "The Computer Model of the Mind", en D. Osherson y E. Smith (eds.), *An invitation to Cognitive Science: Thinking* (Vol. 3). Cambridge, Mass.: MIT Press.

Clarke, D. (2003) *Descartes's Theory of Mind*. Oxford: Clarendon Press.

Clark, A. (2008) *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford: OUP.

Cleland, C. (1993) "Is the Church-Turing thesis true?", *Minds and Machines* **3**: 283-312.

Copeland, J. (2002) "The Church-Turing Thesis", en http://plato.stanford.edu/entries/church-turing/

Cottingham, J. (1998) "Descartes' treatment of animals", en *Descartes*. Oxford: Oxford University Press, pp. 225-233.

Dennett, D. (1991) *Consciousness Explained*. London: Penguin Books.

Dennett, D. y Hofstadter, D. (1981) *The Mind's I: Fantasies and Reflections on Self and Soul*. New York: Basic Books.

Descartes, R. (1996a) *Discours de la Methode,* en C. Adam y P. Tannery (eds.) *Ouvres de Descartes*, Vol VI. Paris: VRIN, pp. 1-78.

_____ (1996b) *Le Monde ou Traité de la Lumiere,* en C. Adam y P. Tannery (eds.) *Ouvres de Descartes*, Vol XI. Paris: VRIN, pp. 3-118.

_____ (1996c) *Principia Philosophiae*, en C. Adam y P. Tannery (eds.) *Ouvres de Descartes*, Vol VIII. Paris: VRIN, pp. 5-353.

_____ (1996d) Meditationes de Prima Philosophia, en C. Adam y P. Tannery (eds.) *Ouvres de Descartes*, Vol VII. Paris: VRIN, pp. 17-90.

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

Heil, J. (2004) *Philosophy of Mind: A Contemporary Introduction*. New York: Routledge.

Kripke, S. (1980) *Naming and Necessity*. Cambridge, Mass.: Harvard University Press.

Lewis, D. (1966) "An argument for the identity theory", *Journal of Philosophy* **63**: 17-25. Reimpreso en J. Heil (ed.), *Philosophy of Mind: A Guide and Anthology*. Oxford: Oxford University Press, pp. 150-157.

Penrose, R. (1999) *The Emperor's New Mind*. Oxford: Oxford University Press.

Putnam, H. (1967) “Psychological predicates” en W. Capitan y D. Merril (eds.), *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press. Reimpreso en J. Heil (ed.), *Philosophy of Mind: A Guide and Anthology*. Oxford: Oxford University Press, pp. 160-167.

_____ (1973) “The nature of mental states”, originalmente publicado como “Psychological Predicates”, en W. Capitan y D. Merril (eds.), *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press. Reimpreso en D. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings.* New York: Oxford University Press, pp. 73-79.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) “Learning internal representations by error propagation”, en D.E. Rumelhart, J.L. McClelland, and the PDP Research group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1. Cambridge, Mass.: MIT Press, pp. 318-62.

Searle, J. (1980), “Minds, brains and programs”, *Behavioral and Brain Sciences* **3**: 417-24. Reimpreso en M. Boden (ed.), *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press, pp. 67-88.

_____ (1990) “Is the brain’s mind a computer program?”, *Scientific American*, January **1990**: 20-25.

_____ (1992) *The Rediscovery of the Mind.* Cambridge, Mass.: MIT Press.

_____ (1995) *The Construction of Social Reality*. London: Penguin.

_____ (2002) “Twenty-one years in the Chinese Room”, en J. Preston y M. Bishop (eds.), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence.* Oxford: Oxford University Press, pp. 51-69.

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

Turing, A. (1936) “On computable numbers, with an application to the *Entscheidungsproblem*”, *Proceedings of the London Mathematical Society*, series 2, Vol. **42**: 231-65 (con correcciones en Vol. **43**: 544-6).

_____ (1950) “Computing intelligence and machinery”, *Mind* **LIX**, no. 2236, Oct. 1950: 433-60. Reimpreso en: M. Boden (ed.) *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press, pp. 40-66.

_____ (1951) “Can Digital Computers Think?”, tipeo de una entrevista radial en el tercer programa de BBC, del 15 de mayo de 1951. Número de referencia de los Archivos Turing: B.5. Reimpreso en S. Shieber (ed.), *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. Cambridge, Mass.: MIT Press, pp. 111-116.