La Tesis de Turing-Church

Esta sesión se va a dedicar íntegramente a discutir algo con lo que, de hecho, ya hemos entrado en contacto. El desarrollo formal de la noción intuitiva de tarea efectiva conduce de forma casi inexorable ante una conjetura de considerables implicaciones: ¿hay alguna tarea o actividad que los seres humanos seamos capaces de realizar de manera efectiva que no caiga bajo el alcance de los formalismos analizados en la sesión anterior? Es importante entender que ésta no es una de esas preguntas que el filósofo acostumbra a hacer para evitarse los sinsabores de la genuina investigación formal. Se trata de una pregunta con un contenido matemático explícito: ¿podemos imaginar procedimientos formales de mayor alcance que aquellos que hemos definido hasta ahora?

La tesis de Church, o de Church-Turing, o aquella que generalmente se expresa mediante cualquier otra combinación de nombres de los protagonistas de ese periodo –excepto, tal vez, Gödel- es aquella que da un respuesta negativa a la pregunta que acabo de mencionar.

[1] *Tesis de Church-Turing*: Toda tarea ejecutable de manera efectiva es computable.

Es decir, no cabría esperar otros procedimientos formales de mayor potencia a los ya conocidos sin violar explícitamente algún componente tangible en la noción intuitiva de tarea efectiva.

Es frecuente encontrar otras formas de exponer esta misma tesis, así como variantes más o menos explícitas. Si la discusión se desarrolla, por ejemplo, dentro de

un estilo más marcadamente formal que el presente, entonces se hablará de funciones numéricas (parciales) definidas sobre los enteros positivos. No obstante, no hay ninguna diferencia. Si la tesis es correcta, entonces no puede haber ninguna tarea efectiva que no sea expresable en términos de operaciones numéricas sobre enteros positivos. En la medida en que cualquier alfabeto no numérico empleado para la descripción de una tarea efectiva puede ser codificado por un procedimiento efectivo oportuno, la diferencia entre palabras y números se desvanece. También es típico utilizar términos más concretos que el de "computabilidad" empleado aquí en [1]. Es decir, en muchas ocasiones en que se discute esta tesis se menciona explícitamente alguno de los formalismos concretos discutidos en la sesión anterior, funciones recursivas generales, o máquinas de Turing, son los más habituales. La tesis que afirma que toda función numérica es recursiva general suele asociarse al nombre de Church, mientras que la tesis según la cual toda función numérica es Turing-computable se relaciona, como es obvio, con el nombre de Turing. En la medida en que ya no hablamos tanto de episodios históricos, sino de una afirmación que recoge el estado actual de la cuestión, parece más apropiado optar por un término genérico capaz dar a entender que no hay diferencia entre cualesquiera de las versiones formales disponibles. Es cierto que en ocasiones se elige alguna de estas opciones con la intención de hacer referencia al estilo o al componente intuitivo que anima a cada uno de estos modelos, pero no veo qué se añade así a la discusión de este principio.

La historia de esta tesis es intensa durante sus primeros años. El primero en hablar de ella, aunque nunca bajo la forma de una tesis, es Church. Según él mismo manifiesta, esto tiene lugar en una conversación con Gödel en la que Church menciona la posibilidad de identificar cálculo-λ con calculabilidad efectiva. La acogida de Gödel es completamente fría: no ve el modo de identificar un concepto informal como el de calculabilidad efectiva con otro técnico y preciso como el de definibilidad en cálculo-λ, el cual, ciertamente, no da mucha idea de la forma en que los aspectos del concepto intuitivo se incorporaran en su *modus operandi*. Church rebate esta objeción mostrándose dispuesto a hablar del propio formalismo de Gödel –recursividad general-

en lugar del suyo. Pero Gödel manifiesta idénticas dudas acerca de esa posibilidad. De hecho, Gödel mantiene esta actitud hasta que el teorema de normalización de Kleene dota al asunto de suficiente generalidad y, sobre todo, hasta la llegada de las máquinas de Turing. Sólo con ellas parece Gödel iniciar su proceso de aceptación de la tesis –proceso que quizá nunca concluyó del todo-.

Tal vez por esta decepcionante acogida, o por la necesidad de adoptar para sus especulaciones las funciones recursivas de Gödel, Church nunca llega a hablar de una tesis. En el abstract de la comunicación que presenta al Congreso de la *American Mathematical Society* de 1935 menciona la posibilidad de *identificar* la noción intuitiva de función numérica efectivamente calculable con la de función recursiva general. En el texto de la comunicación expresa ese mismo hecho bajo la forma de un enunciado, una tesis, que puede ser verdadera o falsa. Es Kleene quien años más tarde se refiere a esta afirmación como una tesis –Tesis I- y le otorga el valor que actualmente se le concede. Turing, por su parte, tampoco formula nunca algo que explícitamente pueda ser reconocido como una tesis. Sin embargo, es cierto que el propio planteamiento de su artículo fundacional permite entender que lo que anda en juego es, exactamente, lo que en [1] se establece.

Desde aquellos años en que todo estaba por hacer se ha escrito mucho sobre esta afirmación. Lo primero que debemos tener en cuenta es que no todos los especialistas coinciden a la hora de evaluar su carácter. Si leemos el enunciado de la tesis nuestras primera impresión es la de encontrarnos ante un enunciado perteneciente al dominio de las ciencias formales. Es un tópico muy extendido considerar que los enunciados característicos de este tipo de disciplinas son de dos tipos: axiomas y teoremas. Los axiomas constituyen verdades evidentes o principios indemostrables por su carácter que tomamos como punto de partida para el desarrollo de una teoría. Los teoremas son verdades demostradas a partir de los axiomas mediante el uso de principios inferenciales previamente admitidos. Parece obvio que la Tesis no se encuentra en ninguna de estas dos circunstancias. No es un axioma en la medida en que acaba por aparecer como consecuencia de una serie de circunstancias previas y más fundamentales que ella desde el punto de vista de la teoría. No es

tampoco un teorema por la sencilla razón de que no existe modo de demostrar su enunciado. ¿Qué es entonces? Si aceptamos el hecho de que constituya un enunciado no trivial, entonces parece que hemos de aceptar también su carácter conjetural. Se trataría, entonces, de una hipótesis empírica –en el sentido popperiano, al menos- dependiente de un dominio formal del conocimiento. Curioso enunciado, ciertamente. Pero esto es así si aceptamos ver en él un enunciado y además uno no trivial.

Hay autores que niegan que [1] sea, propiamente, un enunciado. En su lugar prefieren ver en ello una definición. Es difícil rechazar sin más esta posibilidad porque, como hemos visto, Church mismo sostuvo esta posición, o al menos la adoptó como punto de partida. Como definición indicaría el modo en que debemos entender de manera precisa un concepto informal difícil de manejar. Como tal no sería ni verdadera ni falsa, porque las definiciones no constituyen enunciados. En mi opinión es posible aceptar esta posición sin abandonar la formulación ofrecida en [1]. La razón es que existen muchos tipos de definiciones. Las hay, por ejemplo, *estipulativas*, en las que el concepto definido carece de entidad previa, y las hay *reductivas*, o *aclarativas* en las que el concepto definido existe previamente en un nivel intuitivo. Con éstas últimas se pretende dar una explicación de ese concepto en un dominio que mejora en algo el tratamiento que ante recibía en la intuición informal. Las definiciones de este segundo tipo están asociadas a enunciados, o al menos a condiciones de verdad, de forma obvia: pueden ser adecuadas o inadecuadas y pueden ser completas o incompletas. Aquello a que llamamos tesis de Church-Turing podría ser tomado perfectamente como una expresión de la adecuación y completitud de la definición de “tarea efectiva” por medio del concepto formal de “tarea computable”.

Otra posición que también ha gozado de cierto predicamento es aquella que acepta ver en [1] un enunciado, para otorgarle a continuación un estatus muy peculiar. En concreto, se trataría de una tautología y, por tanto, de un enunciado sin valor desde un punto de vista empírico. Para ver en la tesis de Church-Turing una tautología tenemos que considerar que la noción formal de tarea computable constituye una reformulación más o menos directa de los términos presentes en el concepto intuitivo

de tarea efectiva. Es decir, la descripción que las máquinas de Turing hacen, por ejemplo, de la efectividad no constituye sino una explicación de aquello que el concepto informal significa de hecho para nosotros. Esta posición, así expresada, resulta difícil de atacar. Pienso, no obstante, que oscurece ciertos componentes fundamentales del curso real de los acontecimientos. En primer lugar, nos lleva a pensar que el concepto forma de computabilidad constituye una reformulación directa del concepto intuitivo subyacente. Sin embargo, ya hemos visto que no es así. El concepto intuitivo se ve violentado por su correlato formal en al menos dos sentidos. El concepto informal es, como se dijo en un momento, esencialmente relativo, mientras que el concepto formal es absoluto. Difícilmente aceptaremos que resulte intuitivamente aceptable hablar de efectividad en el caso de una tarea que falla en alcanzar sus objetivos, pese a que la noción formal así lo aconseje. Por otra parte, la negación de algunos componentes del concepto formal no nos sitúa ante una imposibilidad lógica, vista, al menos, desde un punto de vista intuitivo. Ese es el caso, desde luego, cuando aceptamos la existencia de una máquina como la que se propone en el problema de parada, o, como veremos más adelante, cuando imaginamos ciertos procesos extendidos a lo largo del tiempo.

Si me parece, en definitiva, que es mejor dejar las cosas como están y ver en [1] una genuina tesis ello se debe a que la distancia entre el concepto intuitivo y el formal es suficiente como para que quepa imaginar tareas efectivas que no son computables. Es también cierto que en el curso de las numerosas réplicas que esta tesis ha recibido siempre ha sido posible rebatir los ataques objetando que el contraejemplo ofrecido viola en tal o cual aspecto las condiciones impuestas por el concepto intuitivo. No es de extrañar que el abuso de este tipo de crítica haya conducido a ver en [1] una tautología imposible de rebatir con contraejemplos. Reconozco que esta posición es difícil de contestar pero insisto en ver aquí una tesis que puede ser en principio falsada localizando una instancia de tarea intuitivamente efectiva irreproducible en términos de alguna tarea computable.

Ha pasado ya algún tiempo desde que esta tesis alcanzara suficiente fama y se convirtiera en uno de los tópicos de la investigación formal en el siglo xx. El hecho de

que siga siendo ampliamente aceptada por la comunidad científica indica la existencia de un alto grado de confirmación pero también, torciendo considerablemente el argumento, aumenta la probabilidad de una pronta refutación. ¿Qué razones se han aducido para que esta tesis se halle tan sólidamente asentada? Puedo mencionar hasta tres tipos distintos de argumentos generalmente aducidos a su favor.

[2] Argumentos en apoyo de la Tesis de Church-Turing.

i. Inexistencia de contraejemplos aceptables.
ii. Resistencia al efecto de las técnicas de diagonalización.
iii. Existencia de un resultado de normalización: equivalencia extensional.

Como se puede ver fácilmente, estos argumentos son de dos tipos muy distintos. [2.i] se refiere a la inexistencia de refutaciones de la tesis. Es, por tanto, el tipo más fuerte de argumento que cabe dar –al menos desde el punto de vista de las tesis popperianas sobre la refutación y confirmación de teorías-. La posibilidad de concebir contrajemplos que no llegan a confirmarse bastaría para hacer de este enunciado uno con contenido empírico. La polémica sobre el carácter tautológico de la tesis, avivadas siempre que se idea algún contraejemplo, resta valor a lo anterior. Es posible incluir también [2.ii] en este apartado en la medida en que la resistencia a la técnica de diagonalización constituye la superación de una potencial refutación. Gödel siempre vio en este hecho algo casi *milagroso* –este término es literal- que parecía ir en contra de todo lo esperable y de todo aquello que se había confirmado para nociones sumamente próximas. Así, mientras que la noción de derivabilidad, que no es sino un tipo especial de calculabilidad, es relativa a cada sistema y se ve afectada por los efectos negativos de la autorreferencia, esto mismo no se cumple para la computabilidad. Para Gödel este resultado indica que tal vez nos hallemos en presencia de la correcta interpretación formal de un concepto intuitivo, fenómeno extraño, pero no imposible en la historia del pensamiento lógico-matemático.

[2.iii] constituye un género bastante distinto de argumento. Lo que propone es una colección de hipótesis todas las cuales han sido positivamente confirmadas hasta la fecha. El contenido empírico de este argumento es fácil de identificar en este caso ya que conduce directamente a una consecuencia contrastable en la experiencia:

[3] *Contenido empírico del teorema de normalización + Tesis de Church-Turing:*

i. Sea Δ una clase de funciones diseñada con el fin de capturar *sólo* funciones numéricas *efectivamente calculables* –es decir, Δ es correcta en relación al concepto informal de función efectivamente calculable-.

ii. Supongamos que se ha demostrado que la clase Γ formada por todas las funciones Turing-computables guarda con Δ la siguiente relación: $\Gamma \subseteq \Delta$.

Entonces,

iii. Existe una demostración de la afirmación según la cual $\Delta \subseteq \Gamma$.

Esta es una forma sumamente directa de expresar el contenido empírico de la Tesis de Church y de entender por qué goza de la aceptación que todavía posee. Ha habido muchas ocasiones en las que por una u otra razón se ha tenido que construir un nuevo formalismo para representar tareas que es posible ejecutar de manera efectiva. En cualquier momento podemos, si se quiere, intentar definir uno que sea, en la medida de lo posible, distinto de los conocidos. Determinar si un formalismo sólo permite formular tareas efectivas –comprobar si es correcto- constituye un problema

relativamente asequible. El contenido empírico de la tesis de Church-Turing permite conjeturar que lo único preciso para comprobar que es completo –desde su punto de vista- es analizar si puede expresar en su interior las mismas tareas que cualquiera de los procedimientos conocidos hasta ahora. Establecer este extremo es, de nuevo, algo fácil de alcanzar mediante reducciones directas bien conocidas por cualquier lógico. La tesis arroja, desde este punto de vista, predicciones cuya veracidad es posible comprobar por medio de técnicas bien establecidas. Hay muchas hipótesis en el dominio de las ciencias empíricas cuyas condiciones de contrastación son menos claras que las que acompañan a la tesis de Church.

Pese a todo la tesis tiene oponentes, los ha tenido, además, desde siempre. Es fácil confundir y mezclar las objeciones a la tesis de Church con aquellas que se refieren directamente a la identificación de la mente con algún tipo de mecanismo, o con las posiciones conocidas en general bajo el rótulo de *mecanicismo*. Aunque confundir ambos extremos, tesis de Church e hipótesis mecanicista, es prima facie un error, no es uno que sea fácil desterrar. Voy a intentar centrarme primero en aquellas objeciones que tienen como destinatario directo la tesis para hablar a continuación del propio mecanicismo. Las críticas a la Tesis de Church-Turing las ha habido desde ámbitos muy distintos y presentan muy diversas facturas. Me limito a comentar dos que, a mi juicio, aún merodean, unas veces de forma clara, otras de manera implícita, en numerosos debates..

[4] *Objeciones a la Tesis de Church-Turing*

i. Su aceptación posee consecuencias altamente implausibles.

ii. La mente en su actividad es capaz de experimentar un desarrollo ilimitado.

La primera objeción se debe a L. Kalmar y data de 1957. Se trata de un argumento algo sutil, así que habremos de esforzarnos un poco. En primer lugar, consideraremos la función numérica $\Psi(x,y)$ que corresponde a una máquina de tipo universal. $\Psi(x,y)$ es, por tanto, del tipo $\varphi^x(y)$. A partir de $\Psi(x,y)$ podemos definir un predicado numérico como sigue:

[5] T(x,y,z) es verdadero syss $\Psi(x,y)=z$.

El efecto de la diagonalización sobre ese predicado permite establecer la existencia de un entero positivo *n* para el cual sucede que $\neg\exists z\ T(n,n,z)$. Esta consecuencia se sigue de las condiciones particulares bajo las cuales se define la calculabilidad efectiva en términos, por ejemplo, de máquinas. Es decir, la veracidad de $\neg\exists z\ T(n,n,z)$ es una consecuencia de aceptar que las máquinas de Turing computan todas las funciones numéricas efectivamente calculables, pero no disponemos de una demostración directa del enunciado. Nada impide, sin embargo, intentar construir una. La demostración constaría de dos partes. En la primera se tomaría la secuencia T(n,n,1), T(n,n,2), ...y se comprobaría para cada una de ellas si es o no verdadero el enunciado en cuestión. La segunda parte, que procede de manera simultánea a la primera, se propone demostrar la inexistencia de algún entero positivo *k* que satisfaga T(n,n,k) admitiendo para ello cualesquiera medios. Por el contexto de diagonalización y el significado del predicado T(x,y,z), dicho entero no puede existir, y tampoco puede haber una demostración efectiva –interna al sistema del mismo-. Pero Kalmar hace extiende los recursos disponibles a cualesquiera medios, con lo cual nos hallamos ante un enunciado que *sabemos* verdadero pero cuya verdad es *indemostrable* desde cualquier respecto formal. Como veremos en breve, este tema apunta a uno de los tópicos dentro del debate mente-máquina. En esta ocasión tiene un sabor estrictamente formal, pero apunta ya al valor que deben tener aquellas demostraciones que establecen limitaciones acerca de la potencia de un formalismo. Kleene, mucho menos inclinado a la filosofía que Kalmar, zanjó la

cuestión indicando que se trata de una mala comprensión de los resultados de limitación implicados por la técnica de diagonalización y que Kalmar, en cualquier caso, no ofrecía en ningún momento la definición explícita de una función –predicado de números- al mencionar la admisión de medios probatorios que no están en principio fijados.

Y es que el carácter finito de la descripción de una tarea efectiva ha sido siempre para Kleene uno de los rasgos definitorios de la computabilidad. La segunda objeción a la tesis, debida esta vez al propio Gödel, apunta en esa dirección. Gödel admite que la mente humana pueda comportarse en cada instante como una máquina de Turing, de tal modo que las tareas que es capaz de ejecutar de manera efectiva nunca podrían superar aquellas que una máquina de Turing puede igualmente ejecutar. Sin embargo, no hay nada que garantice que algunas de las tareas que los seres humanos llevamos a cabo de manera efectiva no se realicen siguiendo una rutina íntimamente ligada a la evolución de los estados de nuestra mente, evolución que, según Gödel, podría *tender a infinito*. Esta observación ha sido considerada siempre como algo sumamente enigmático que merece atención debido en parte a la autoridad de quien lo afirma. Kleene despacha el argumento en similares términos declarándolo un sinsentido. Si la descripción de un algoritmo contiene algún elemento abierto o indeterminado que se puede manifestar en rasgos imprevistos de una posible evolución del propio sistema estaremos, afirma Kleene, ante algo que no merece el nombre de algoritmo. Se habrán violado simplemente, las condiciones de contorno del problema. En este caso lo que parece estar en juego es el modo en que se puede combinar el carácter universal de la computación con la evolución del propio algoritmo. Es fácil pensar que aquello a lo que Gödel se refiere es a la evidente capacidad que la mente humana posee para evolucionar cambiando la rutina que aplica a la resolución de uno y el mismo problema. En la medida en que una máquina es una entidad perfectamente definida desde un principio, este tipo de plasticidad le estaría prohibida conduciendo así a la constatación de tareas que nosotros ejecutamos efectivamente sin especial esfuerzo pero que quedan fuera del alcance de una máquina. Sin embargo, no es la plasticidad o rigidez lo que se discute: una máquina universal posee la suficiente capacidad como para cambiar la rutina que ejecuta tantas veces como

quiera en el curso de un cómputo. Lo único que es preciso suponer es que el mecanismo de control que determina en qué momento y a partir de qué consideraciones debe ser alterada la rutina en curso es también un elemento que puede ser descrito finitamente y de una vez por todas. Son varios los autores que han apuntado a que el único modo de entender la objeción de Gödel es negando este extremo. Si tomamos en serio su sugerencia, nuestra mente sería capaz de ejecutar tareas mediante rutinas que en cada instante pueden ser descritas como máquinas de Turing pero cuya serie total –cada intervalo finito también lo sería- no puede serlo. ¿Hay alguna tarea cuya efectividad dependa de esta conjetura? Esto está por probar.

La cuestión que el mecanicismo plantea es la de si es posible identificar todas o parte de las capacidades del ser humano con aquellas que se puede incorporar en algún ingenio mecánico fruto de la propia mano del hombre. Generalmente se ha entendido que las tesis mecanicistas no sólo analizan este problema, sino que además pretenden ofrecer una respuesta afirmativa al mismo. En la actualidad no disponemos de datos capaces de inclinar la balanza en una u otra dirección y cabe pensar que tal vez nunca lleguemos a resolver la cuestión por completo, al menos en su forma actual.

Resulta difícil encontrar un solo momento en la historia del pensamiento occidental en el que no se haya planteado este problema en alguna de sus formas. La actualidad de la cual goza en el presente se debe, en parte, a que esa forma ha cambiado dando lugar a nuevos planteamientos. El estudio de la posibilidad de reproducir habilidades característicamente humanas en ingenios artificiales depende, como no podía ser de otra forma, del tipo de medio o soporte sobre el cual se cree posible incorporar esas facultades. El descubrimiento de una nueva teoría o de nuevos recursos suele provocar un avance sensible en los planteamientos del mecanicismo. El modelo vigente desde el triunfo de la nueva ciencia durante los siglos xvii y xviii era el de la mecánica racional. La imitación de las facultades humanas se veía y estudiaba en el contexto que esa teoría brindaba. La mecánica racional suministra un modelo caracterizado fuertemente por la previsibilidad de los estados que el sistema puede experimentar sorprendiendo, ante todo, la posibilidad de reproducir por medios mecánicos el movimiento natural. Otra de las características de este modelo es su

conducta local: cada ingenio parece tener una función específica dotada de muy poca plasticidad. La excepción a la regla lo ofrece la *analitical engine* de Babbage, mucho más próxima a un ordenador moderno que ingenios posteriores técnicamente mucho más capaces.

Es obvio que éste no es ya el paradigma en torno al cual se desarrolla el debate mecanicista. La razón de ello es, en buena medida, la rendija que la tesis de Church-Turing abre y por la cual penetra con fuerza todo el caudal de ideas asociadas a la ciencia cognitiva y la inteligencia artificial. Para entender esta afirmación bastará pensar por un momento qué hubiera pasado si hubiéramos dispuesto de un buen contraejemplo de esta tesis. Deberíamos aceptar la existencia de tareas efectivas, tal vez no bien conocidas, pero sí al menos posibles en principio, que el ser humano concibe como tales y que no pueden ser descritas en términos de ningún formalismo particular. Nos encontraríamos así con que los ingenios mecánicos basados en esos formalismos nunca podrían ser empleados como modelos fidedignos de la mente humana, ni siquiera en aquello que podemos considerar más básico: la interpretación de tareas efectivas. La tesis de Church-Turing, más que decir algo manifiestamente a favor del mecanicismo permite, simplemente, que la Teoría de la Computación reemplace a la mecánica racional como marco teórico para el mecanicismo. ¿Qué rasgos de la Teoría de la Computación son los que más claramente determinan ese nuevo modelo? Al tratarse de una disciplina formal cuyos resultados están fuertemente conectados unos con otros por medio de relaciones lógicas elementales, ésta puede ser una cuestión difícil de responder. No obstante, y a un nivel puramente intuitivo, sí parece que lo siguiente debe ser tenido en cuenta:

[6] *Modelo mecanicista computacional*:

i. Es un modelo formal.
ii. Contiene instancias dotadas de la máxima plasticidad.
iii. Posee limitaciones dictadas por el propio formalismo.

Es un modelo formal en la medida en que el soporte material sobre el que se implementa es totalmente independiente de la función ejecutada. La plasticidad la debemos a la presencia de suficientes dosis de autorreferencia –fenómeno privativo de los sistemas sígnicos superiores- que permite, mediante operaciones sobre códigos, que nuestros modelos experimenten alteraciones, evolucionen, etc. Todos estos cambios pueden tener lugar sin perder en todo momento la identidad del objeto que los experimenta. Una máquina universal de Turing puede ejecutar cualquier tarea concreta sin perder su propia identidad como rutina. Por último, no nos encontramos esta vez con un modelo teórico en el que el carácter mecánico de una actividad equivalga a la posesión de un pleno control sobre su conducta. Los resultados de limitación estudiados en sesiones anteriores suministran esta vez una concepción mucho más matizada, y desde luego mucho más modesta de lo que significa proceder de manera mecánica.

El principal argumento a favor de la tesis mecanicista, interpretada bajo el nuevo modelo computacional, lo aporta Turing en un ensayo considerado hoy un clásico. El título con el que ha llegado a ser conocido, *¿Puede pensar una Máquina?*, es suficientemente elocuente. En él se propone un juego, denominado *juego de la imitación*, en el que con una sencillez demoledora se establecen las condiciones bajo las cuales cabe atribuir pensamiento a una máquina cuya programación se ajuste a los principios establecidos por la Teoría de la Computación. Este juego se plantea como un test –*Test de Turing*- en el que una máquina convenientemente programada es sometida a una batería de preguntas mediante las cuales se intenta establecer si la conversación tiene lugar con un ser humano o con un ingenio mecánico –un ordenador En el original el juego se propone inicialmente para intentar averiguar si se está ante un hombre o una mujer-. Como es obvio, se han eliminado todos los rastros que pueden llevar al experimentador a descubrir quién es su interlocutor debido a diferencias físicas apreciables. Se trata de un experimento mental en el que las condiciones se pueden extremar a placer. Turing propone que en caso de ser incapaces de identificar un ingenio mecánico aceptemos como un hecho bruto hallarnos ante un ser dotado de inteligencia, aunque ésta no sea natural. Anticipa, eso sí, sus dudas sobre la existencia de algún mecanismo capaz de superar este test y

hace un pronóstico: en 50 años a partir de la fecha en que se publica su ensayo cree posible contar ya con los primeros seres capaces de superarlo. Si tenemos en cuenta que este artículo está fechado en 1950, la cuestión no puede ser más actual.

Si prescindimos del entorno colorista con que generalmente se lee este ensayo, no nos será muy difícil reconocer que estamos ante una pieza fundamental del pensamiento contemporáneo. Creo, además, que el término *tesis de Turing* podría emplearse con suma propiedad para hacer referencia a la siguiente afirmación:

[7] *Tesis de Turing*:

- Si bajo las condiciones más estrictas y generales que quepa imaginar es imposible descubrir a partir de la sola conducta de una entidad si su origen es artificial o natural, entonces los mismos predicados que tendría, si supiésemos que es natural, deben serle aplicados, aun cuando podamos descubrir por otros medios –independientes de su conducta- que no lo es.

Se trata de un enunciado difícil de formular pero fácil de entender. Si no hay razones que permitan discriminar el comportamiento de dos entidades, aún cuando puedan tener soportes materiales muy distintos, entonces se les debe atribuir los mismos predicados, las mismas propiedades, en definitiva. Y esto vale para la inteligencia, la conciencia e incluso, por qué no, para los derechos civiles de esas entidades. Además, en caso de descubrir que el soporte material es distinto por un medio directo, nunca apreciable en la conducta, tampoco habría razones para modificar los predicados relevantes que antes se atribuían a esa entidad. En definitiva, es la función la que prima sobre la materia a la hora de fijar las propiedades relevantes que caracterizan un ser.

Este principio tiene una larga tradición que se remonta a la persecución que Ockham hace de entidades inobservables capaces de determinar la *quiditas* de cada entidad particular o de cada género. El problema retorna ahora centrado en la atribución de inteligencia a seres mecánicos. ¿Qué esencia o que fantasma es ese, imposible de determinar por cualesquiera medios, que garantiza que mis actos mentales poseen una entidad que nunca podrá tener un ingenio mecánico cuya conducta sea en todo lo demás humana? ¿Por qué creo yo que mi conducta responde a mis actos mentales mientras que en el caso de un ingenio mecánico sólo responde a una muda secuencia de transformaciones simbólicas ejecutadas de acuerdo a un programa? ¿Cómo es posible negar la tesis de Turing sin caer de una forma u otra en vicios filosóficos tan bien conocidos como el del solipsismo?

Parece obvio que la Inteligencia Artificial, tomada, si es que es posible, como una disciplina científica cohesionada, vería en la tesis de Turing –el enunciado [7]- su principal axioma.

La tesis de Turing expuesta en [7] constituye lo que en la tradición filosófica corresponde a un *sistema máximo* - término empleado por Galileo para referirse al enfrentamiento total existente entre dos concepciones del universo en sí mismas completas: la copernicana y la ptolemaica-.

Se trata de una interpretación de la realidad dotada de una considerable coherencia interna y cierta en apariencia pero enfrentada a algún otro sistema similar, no menos coherente y cierto, pero incompatible con el primero. ¿Qué es lo que impide que aceptemos la tesis de Turing y con ella el paradigma mecanicista? Searle, en un artículo igualmente famoso, establece un experimento mental considerado por muchos el único, o uno de los pocos, que es capaz de enfrentarse y mantener la tensión con el juego de la imitación de Turing. Me refiero al experimento conocido con el nombre de *la habitación china*. El argumento es, como casi siempre que algo importante anda en juego, sumamente fácil de entender. En una habitación se encuentra una persona a la que por una apertura se le suministran ideogramas chinos que puede identificar en un listado y que reemplaza por aquellos signos que figuran asociados a cada ideograma

en ese listado. Una vez completado el reemplazo devuelve el papel por la rendija. Aparentemente esa persona sabe chino, ha traducido una secuencia de símbolos escritos en ese idioma. Sin embargo, ella misma es conscientemente de estar realizando una tarea de tipo simbólico carente por completo de significado. Las condiciones se pueden hacer, de nuevo, tan extremas como se quiera. Es conveniente considerar, por ejemplo, que el pretendido traductor traduce del chino a un idioma que tampoco es el suyo. Lo que este argumento pretende mostrar se puede exponer también en forma de una tesis:

[8] *Tesis de Searle*:

> - Existe una diferencia esencial entre el acto de *reproducir un algoritmo* y *encarnar un algoritmo* que hace que en el primer caso la tarea carezca de significado mientras que en el segundo posea uno intrínseco impidiendo que se apliquen en ambos casos los mismos predicados.

Este enunciado no constituye, por sí sólo, una réplica a las tesis mecanicistas; no niega que seamos máquinas de un cierto tipo, pero rechaza que esa conjetura pueda ser identificada con el hecho de ser máquinas simbólicas del tipo al que corresponden las máquinas de Turing.

Estamos, ya lo he dicho, ante dos sistemas máximos capaces de mantener una tensión a m juicio imposible de dirimir por el momento. Ambos se conceden todo lo que se pueden conceder mutuamente para negar a continuación las principales conclusiones del otro. Cuando en ciencia se alcanza un punto de enfrentamiento tal suele hacer falta algún elemento nuevo por completo que afecte al equilibrio alcanzado, y no hay nada, por el momento, que pueda realizar esa función.

Es importante entender que la respuesta de Searle podría partir de conceder la existencia de un ingenio capaz de superar el test de Turing. Por eso mismo he querido oponer una posición a la otra en este nivel máximo enfrentamiento. Pero no existe hasta la fecha tal ingenio. A continuación y para terminar, voy a repasar una serie de objeciones que se dirigen a mostrar que nunca podrá darse tal circunstancia. Para distinguir estas críticas de las tesis anteriores las consideraré como argumentos.

[9] *Argumento acerca de los resultados de limitación:*

- Ningún algoritmo puede establecer resultados acerca de sus propias limitaciones.

He aquí una razón para pensar que el ser humano siempre será capaz de afrontar empresas que nunca podrán ser abordadas por máquinas de Turing y que, curiosamente, procede del propio contexto en que éstas se definen. Este argumento ha sido popularizado por el filósofo americano John Lucas en torno a 1964 llegando a alcanzar cierto predicamento. Se basa en el impacto que en general producen resultados como los teoremas de incompletitud de Gödel o el propio problema de parada en personas que no conocen el detalle concreto de estos teoremas. Pero con esto no pretendo trivializar el argumento de Lucas. Lo que sí conviene indicar es que su presentación suele resultar en la mayoría de los casos falaz. Se olvida, y aquí está el error, que tales resultados constituyen instancias de demostraciones que pueden ser reducidas a un programa perfectamente capaz de generarlas. Los programas de demostración automática que han conseguido obtener resultados matemáticos tan poco triviales como los teoremas de Gödel son conocidos de antiguo. Construir una máquina de Turing que demuestre el problema de parada puede quedar casi como ejercicio a partir de la demostración que aquí he dado de ese teorema. Toda explicación que permita construir un argumento a favor o en contra de una tesis puede, desde la perspectiva del mecanicismo contemporáneo, dar lugar a una aclaración formal transformable, en última instancia, en un programa.

El argumento de Lucas tiene, no obstante, elementos de mayor mérito. Por un lado, sugiere la relativa incapacidad que un toda entidad mecánica tendría para autoinvocarse mediante alguna forma de conciencia. Esa incapacidad le privaría de construir enunciados en los que ella misma actúa como sujeto de propiedades que se dispone a establecer. El carácter lingüístico y autorreferencial del nuevo modelo mecanicista echa por tierra esta lectura del argumento. Pero hay otro componente que, finalmente, sí aporta algo substantivo. Todas las razones que el mecanicista ofrece para negar la originalidad y carácter privativo de alguna capacidad humana se basan en la construcción de un algoritmo a partir de una descripción objetiva de esas mismas capacidades. Pero una cosa es producir una copia mecánica de un argumento informal, aunque convincente, y otra muy distinta producir un argumento informal de esas características. En otras palabras, podemos aceptar que cualquier resultado de limitación es, en última instancia, una demostración computable que puede ser *reproducida* por una máquina de Turing, pero ¿puede ser *producido* por una de ellas? La cuestión tiene que ver con la espontaneidad más que con el carácter efectivo de una demostración. Hay razones para pensar que algunos de los próximos episodios del mecanicismo van a jugarse en el terreno marcado por la diferencia entre producir un argumento y reproducir uno ya dado. Existen razones algo más profundas que indican desajustes en el modo en que la Teoría de la Computación entiende y relaciona estos conceptos, pero no entraremos ellos. No obstante, bien puede ser que sólo estemos ante un problema provocado por la falta de experiencia en el tratamiento de programas suficientemente sofisticados. Más en concreto, con programas dotados de cierta autonomía y plasticidad.

Este es, precisamente, el objeto del siguiente argumento.

[10] *Argumento sobre la plasticidad*:

- Ningún modelo computacional puede mostrar la misma plasticidad y versatilidad que el ser humano presenta, siendo capaz de evolucionar de forma muy notable sin perder su identidad.

Este es un argumento que, según quien lo defienda, puede indicar ignorancia o una considerable sutileza. Ignorancia si con ello se pretende aludir al carácter concreto, limitado, de una sola función, de los ingenios mecánicos. Éste es un rasgo que la existencia de máquinas universales relega al pasado, al marco anterior presidido por la mecánica racional. Las rutinas pueden experimentar evoluciones, cambios, procesos de aprendizaje, movimientos destinados a la corrección de errores a aprender de ellos, similares por entero a los que el ser humano experimenta. Y sutileza si pretende explotar la crítica que Gödel hace a la Tesis de Church-Turing. ¿Existen procesos que los seres humanos seguimos en ausencia de un mecanismo último de control, los cuales son no obstante, localmente computables? ¿Estamos los seres humanos dispuestos ante la flecha del tiempo del mismo modo que los mecanismos simbólicos cuando ejecutan su programa?

La última objeción procede también de Gödel y es, de todas, la más peculiar.

[11] *Argumento de los actos mentales no simbólicos*:

> -Hay actos mentales que no pueden, por su carácter, ser representados por medio de ningún sistema simbólico o que no dependen de ningún soporte material imaginable.

Este argumento ha suscitado cierta atención a raíz de la última recopilación que Hao Wang, comentarista y discípulo de la obra de Gödel, hace de las opiniones filosóficas de éste último. El referente de Gödel es, en mi opinión, la Tesis de Church y las posibilidades puramente lógicas de refutarla. Una de ellas es, como resulta obvio, admitir la existencia de pensamiento que no pueda ser representado simbólicamente. Dentro de esta categoría caen fenómenos muy distintos. Por una parte, el denominado pensamiento no lingüístico, si tal cosa existe. En la medida en que pueda darse una actividad cognitiva superior con consecuencias causales sobre otros actos mentales que escape a cualquier representación simbólica, es evidente que ésta también quedará fuera del alcance de los formalismos diseñados para caracterizar la

efectividad. También se puede incluir en este apartado la ocurrencia de actos mentales sin partes, o actos mentales instantáneos asociados a la manifestación pura de la voluntad. Aunque parezca algo completamente extravagante, es posible aducir este tipo de fenómeno para explicar ciertos actos de la voluntad que podrían dar razón de por qué los seres vivos no parecemos aquejados del tipo de inacción que el problema de parada pronostica. Hay otras muchas formas de explicar el fenómeno, desde luego, pero también es posible diseñar experimentos mentales inquietantes capaces de situarnos, si realmente somos máquinas simbólicas de considerable complejidad, muy cerca de bucles o estados retroalimentados que nunca tienen finalmente lugar. La posibilidad de que el pensamiento pueda tener lugar con independencia de soporte material tangible es una posibilidad que Gödel toma muy en serio movido en parte por sus opiniones en materia religiosa. Podemos secularizar el argumento apuntando a una nueva vía de investigación abierta recientemente: la computación cuántica. No digo que en ella no concurra un substrato material, sino que en este caso es el tratamiento del tiempo –superposición cuántica- el que alteraría notablemente nuestras consideraciones. Si nuestra mente es, en su ejercicio, sensible, como se ha sugerido en ocasiones, a efectos cuánticos las consecuencias pueden ser imprevisibles ya que escaparíamos del nivel fenoménico admitiendo efectos causales alimentados por una lógica no-clásica.

Orientación bibliográfica.

Hay una obra de la que no he hablado hasta ahora y que puede ser considerada en este momento. Se trata de **[Webb, 1980]**. Esta extensa monografía está dedicada en un sentido general al mecanicismo pero desde los resultados de la ciencia moderna. En el cap. IV se trata la Tesis de Church con cierta profundidad aunque la argumentación resulte a veces confusa. Es una referencia crítica ya clásica por lo sugerente. Su lectura es compleja pero con un poco de ayuda merece la pena.

Para analizar con detalle la réplica de Kalmar a la Tesis de Church, **[Kalmar, 1957]**. Como curiosidad, podemos mencionar una monografía sobre la Tesis de Church en lengua portuguesa, bastante completa aunque no muy profunda, **[Estola Biraben, 1996]**.

Otro librito que conviene tener a mano es **[Turing, Putnam, y Davidson, 1985]** ya que además contiene **[Turing, 1950]**. El artículo de Searle donde se responden la tesis de este último texto es **[Searle, 1980]**, otro auténtico clásico. Para seguir las complejas opiniones de Gödel y sus sugerencias disponemos de **[Wang, 1974]** y **[Wang, 1996]**. La crítica de Lucas procede de **[Lucas, 1961]** y aparecen también recogidas en **[Penrose, 1989]**, cap. 4 y en **[Webb, 1980]**, cap. 4. **[Martínez-Freire, 1995]**, cap. 8 y . **[Martínez-Freire, 2000]** son buenas bases para una discusión de carácter filosófico.

Como referencias más bien orientadas a la historia de los acontecimientos cabe citar los que ya se han mencionado en otras ocasiones, es decir, **[Kleene, 1981]**, **[Kleene, 1987]**, **[Kleene, 1994]**, **[Rosser, 1984]**, **[Davis, 1982]** y **[Sieg, 1997]**. Un texto interesante por su peculiar interpretación es **[Gandy, 1994]**. **[Penrose, 1989]** cap. 2 es demasiado breve.

A parte de esto, casi cualquier manual trae referencias al respecto. Destaco **[Kleene, 1952]** secc.62 por ser este autor quien dio nombre a esta tesis, y quien más de cerca vivió su desarrollo. **[Boolos, 1974]**, cap. 6, **[Salomaa, 1985]** cap. 4, etc. contienen todos ellos comentarios al respecto a propósito de sus exposiciones formales.