Cuadernos del Sur. Filosofía
versión On-line ISSN 2362-2989

Cuad. Sur, Filos. no.39 Bahía Blanca 2010

Dejando de jugar al juego de la imitación

Jorge Mux[*]

[*] Licenciado en Filosofía. Profesor adjunto de las cátedras *Filosofía del Lenguaje* y *Problemas de la Filosofía* en la Universidad Nacional del Sur.

Resumen
El clásico experimento mental de Alan Turingacerca de la indistinguibilidad entre una máquina y un humano pretende poner a prueba la capacidad de respuesta de una máquina de estados discretos, pero no queda claro qué puede esperarse a partir de las respuestas dadas por la máquina, dado que las clases de contraste entre las alternativas de respuesta que suscita la prueba no son lo suficientemente marcadas para establecer algún tipo de decisión epistémica. Por otra parte, en el test resulta crucial la capacidad de hacer preguntas del interrogador y la capacidad de engaño de la máquina, elementos que agregan un grado de arbitrariedad a la prueba. En su lugar se propone un test en el cual no se evaluará la capacidad de pensamiento de una máquina sino la posibilidad de que dicha máquina sea un modelo -parcial o total- no trivial del pensamiento.

Palabras clave: Mente: Máquina; Turing.

Abstract
Alan Turing's classical thought experiment on the indistinguishability between a machine and a human being relies upon testing a discrete state machine's ability to reply questions. However, what to expect from the answers given by the machine during the experiment remains unclear, as the ways to contrast those replies are not well defined enough to serve as a basis for any epistemic decision. Moreover, the test is crucially dependent on the interrogator's ability to ask questions and the machine's ability to deceive the experiment's judge. These elements add some level of arbitrariness to the test. We propose instead a test which will not try to assess a machine's ability to think, but the possibility of that machine's being a (non trivial) partial or total model of thought.

Key words: Mind; Machine; Turing.

Fecha de recepción: 24 de Agosto de 2010
Aceptado para su publicación: 26 de Agosto de 2011

Cuando Alan Turing propone el clásico juego de la imitación, no pretende responder a la pregunta "*¿Pueden pensar las máquinas*?" porque la considera carente de sentido. En su lugar, propone la pregunta "*¿Qué sucede cuando una máquina juega el juego de la imitación*?". Al hablar de "máquina", Turing excluye expresamente a los seres biológicos. El resultado del test pone de relieve que la *capacidad pararesponder* y *engañar* son los elementos cruciales para lograr una indistinguibilidad entre computadora y humano.

Turing presenta en primer lugar un problema -el de la posibilidad de que las máquinas piensen- y luego sustituye ese problema por otro: "¿qué sucede cuando una máquina juega al juego de la imitación?". Es muy común en filosofía la sustitución del planteo de un problema por otro, con la esperanza de que las respuestas al nuevo problema puedan ser más claras y precisas que las del primero. Sin embargo,¿es lícito sustituir un problema de naturaleza conceptual por otro puramente operacional? ¿En qué sentido esta sustitución nos da respuestas importantes? La estrategia de Turing consiste en calificar de "carente de sentido" al problema inicial, dado que el término "pensar" resulta demasiado impreciso.La noción de "sinsentido" está en consonancia con el positivismo lógico y podríamos parafrasearla del siguiente modo: "no es posible encontrar un método de verificación (para la proposición o el término en cuestión)". Aparentemente, la pregunta "¿pueden pensar las máquinas?" nos induce a una "actitud peligrosa", dado que nos embarcaríamos en las acepciones corrientes de "pensar" y de "máquina" (Turing, 1984:11). Mientras el término "pensar", en su acepción corriente, parece ser "peligroso" y nos lleva a una formulación "carente de sentido", el término

"máquina" puede ser definido con precisión: una máquina es una computadora digital de estados discretos. Sin embargo, ¿es correcto decir que no tenemos un "método de verificación" para el término "pensar"? El propio positivismo lógico, de la mano del primer Wittgenstein, se ocupó de este término y de su especial relación con el lenguaje -factores ambos decisivos para la operatividad del test de Turing-. Para el positivismo lógico es absurdo considerar un pensamiento que no posea soporte linguístico o, dicho de otro modo, un pensamiento que no sea, por principio, comunicable. Los positivistas lógicos no niegan que en la mente humana hay elementos incomunicables, pero éstos quedan relegados a la categoría de "representaciones" o "ideas" (*Vorstellungen*). De esta manera, reducido convenientemente el rango del concepto "pensamiento", podemos decir que el pensamiento equivale a lo que es linguísticamente comunicable: el pensamiento es un proceso mental cuyo correlato objetivo es el lenguaje. Así, para el positivismo lógico solo es pensamiento aquello que puede ser objeto del estudio de la lógica, y con esta identificación consigue que la psicología no pueda interferir en su objeto.

Ahora bien, ¿es toda expresión linguística la expresión de un pensamiento? Por supuesto que no: solo expresan pensamientos aquellas porciones del lenguaje que afirman algo sobre el mundo: aquellas expresiones de las cuales podemos predicar verdad o falsedad. Una pregunta es, en el mejor de los casos, un pensamiento incompleto, cuya completud estará dada por la pregunta más su respectiva respuesta. De este modo, el positivismo lógico realiza una reducción de ambos conceptos de manera tal que uno es la contracara del otro. El problema del origen del lenguaje y del pensamiento quedan reducidos a un solo problema, y resulta absurdo preguntar si el lenguaje es anterior al pensamiento o viceversa, puesto que ambos son idénticos, excepto por el hecho de que el lenguaje es el soporte material del pensamiento. Desde esta perspectiva verificacionista, es evidente que la pregunta "¿pueden pensar las máquinas?" no es un sinsentido, dado que ambos términos clave de la pregunta pueden definirse con precisión[1]. Turing no necesitaba trocar un problema por otro, al menos no por las razones que él aduce con respecto al "sinsentido" del primer problema: parece perfectamente posible asignarle sentido a esa pregunta inicial.

En segundo lugar, ¿qué se gana con modificar la pregunta inicial? O, en otras palabras, ¿por qué la segunda pregunta, *"¿qué sucede cuando una máquina juega al juego de la imitación?*" es mejor que la pregunta inicial? Sin duda, esta pregunta tiene la ventaja de que, dado el experimento propuesto, es fácil responder lo que sucedería en cada caso. Si somos engañados de manera sistemática por un ordenador digital, quizás podríamos decir que el ordenador se comporta de manera casi humana, o que aparenta un enorme grado de inteligencia y vivacidad. Un problema central con la formulación de esta pregunta es que no queda claro cuál es la *clase de contraste* (Van Fraassen, 1977:1143-1150) que se espera con ella. Una clase de contraste es una clase de alternativas frente a las que se contrapone el hecho cuya explicación se requiere. En cambio, con la pregunta original, las clases de contraste pueden quedar mucho más marcadas y de ese modo puede acotarse mejor el aspecto crucial que esperamos obtener de la observación. Pongamos ejemplos de diferentes clases de contraste para la primera pregunta:

CASO A

1. ¿Por qué dices que Carlos es(tá) consciente? (En contraste a "Carlos está desmayado")

2. ¿Por qué dices que la computadora es consciente? (En contraste a "No tiene consciencia en absoluto, como tampoco posee inconsciencia")

Como se ve aquí, cuando atribuimos pensamiento a una máquina no queremos decir, con ello, que las máquinas no pensantes son máquinas desmayadas o algo por el estilo. El contraste que queremos marcar queda claro, y resulta relevante que nuestra pregunta haga precisamente esa distinción.

Ahora pongamos ejemplos de contraste para la segunda pregunta:

CASO B

1. ¿Por qué dices que Carlos ha jugado al juego de la imitación? (En contraste con "Carlos no ha jugado al juego de la imitación")

2. ¿Por qué dices que esta máquina ha jugado al juego de la imitación? (En contraste a "Esta máquina no ha jugado al juego de la imitación")

Como se puede ver, en el caso B las relaciones de contraste son poco marcadas y no hacemos una gran ganancia conceptual si contestamos ambas preguntas. En cambio, en el caso A, es claro que con el término

"pensar" queremos significar algo nuevo; queremos establecer un nuevo tipo de contraste entre entidades con propiedades heterogéneas. En ese sentido, la primera pregunta parece mejor que la segunda.

Por otra parte, ¿qué se puede decir del tipo de respuestas esperables para la segunda pregunta? Más arriba apuntábamos a la posibilidad de que alguien pueda decir que la máquina piensa, o que tiene una vivacidad casi humana. Es aquí donde parece descansar el carácter crucial del juego de la imitación: en *lo que se diga* de la máquina *después* del test. He aquí que, si el humano ha sido engañado por la máquina, es posible que su conclusión sea "estoy frente a un ser humano". Pero, ¿cómo podemos establecer las condiciones por las cuales una persona puede ser engañada? Y, más importante aun, ¿podemos sacar alguna conclusión *acerca de la máquina*, a partir del engaño en que hemos caído? Por un lado, es fácil que una máquina nos engañe, tal como lo han demostrado experimentos con programas del tipo ELIZA o PARRY: el interlocutor puede no sospechar que está hablando con un programa, pero no por una especial capacidad de engaño de dicha máquina, sino porque no tiene motivos iniciales para creer en algo extraño. Cuando algunos psicólogos fueron llevados a dialogar con el paranoico PARRY, en algunos casos diagnosticaron que PARRY tenía una severa lesión cerebral, ¡pero en ningún momento pensaron que aquello con lo cual dialogaban no era humano! (Boden, 1984:143). Esto podría mostrar que aun en casos límite, podemos creer que una actitud es humana sin que efectivamente lo sea. Por otra parte, el sujeto que realiza las preguntas puede ser alguien con escaso conocimiento del mundo, tímido, introvertido e incluso autista. Tampoco debemos olvidar que en este experimento juegan *dos* seres humanos. ¿Qué decir de la otra persona, la que compite con la computadora? Podría equivocarse al transmitir por el teletipo, o podría desconocer cuestiones elementales, o puede tener un corpus de conocimiento diferente debido a que proviene de una extracción social distinta de quien pregunta, o puede ser extranjero. El experimento parece diseñado para que sea realizado por personas de extracciones sociales semejantes, con cierta homogeneidad en su educación y cierta fluidez en el diálogo. Por ello, cuenta con una cantidad infinita de cláusulas *ceteris paribus*, que dependen de quien hace las preguntas y del sujeto humano que está al otro lado de la terminal y que juega a "parecer humano". Estas cláusulas vuelven a la "capacidad de engañar" de la máquina virtualmente irrelevante.

¿Por qué es problemática la pregunta "pueden pensar las máquinas"?

Probablemente el calificativo de *sinsentido* se debe a que "¿Pueden pensar las máquinas?" es un tipo de pregunta *desviada*. Una expresión es desviada cuando se la utiliza en un contexto inusual, proyectando la intensionalidad del término en un uso no común. Dado que "pensar" es por definición una propiedad de seres biológicos, hacer la pregunta para algo que no sea biológico no parece tener pertinencia: cuando realizamos el análisis componencial del término "pensar" puede saltar a la vista que ninguna de las unidades léxicas del análisis de "pensar" nos remite al concepto "máquina". Sin embargo, una proposición "desviada" puede convertirse en "no desviada" si se encuentran contextos de uso que amplíen la referencia (Putnam, 1984:113). El término "navegar", por ejemplo, solo tenía aplicaciones navales hasta que se lo utilizó en computación: ganó un contexto de uso sin perder su contexto original. Sin embargo, existe una diferencia entre el término "navegar" y el término "pensar". Quien dice "estoy navegando en internet" no pretende que se entienda el término "navegar" como un proceso funcionalmente idéntico al de la navegación por mar o por río. En cambio, cuando preguntamos si las máquinas pueden pensar no solo proyectamos un nuevo contexto de uso; también pretendemos que en ese contexto de uso la intensión del término se conserve tal como en su contexto original. No queremos preguntar si la máquina hace *como si* pensara: queremos saber si podemos atribuirle con propiedad todo aquello que le atribuimos a los seres que piensan. Pero aquí la indeterminación es aun mayor: ¿tenemos en claro qué debe atribuirse a un ser para que lo califiquemos de "pensante"? Si dudamos acerca de la capacidad de pensamiento de una computadora, ¿por qué no dudar, también, de la capacidad pensante de otros seres humanos? No conocemos aun cuáles son las condiciones necesarias y suficientes para atribuir pensamiento a algo, y este parece un buen argumento para calificar de sinsentido a cualquier programa que se proponga preguntar si tal o cual cosa piensa.

Por otra parte, si aceptamos que una determinada máquina puede pensar, corremos el peligro de apresurarnos. Quizás lo que la máquina haga sea un sustituto funcional muy grosero del pensamiento. Podríamos decir que la máquina *entiende*, que *deduce*, que *habla*, pero a condición de que todos estos términos se usen de un modo que no reflejen toda la riqueza semántica del *auténtico* pensar.

Otro problema que se suscita con el término *pensar* es que, cuando se determina que algo no puede pensar, no queda claro hasta qué punto se trata de una imposibilidad lógica o puramente empírica. Para algunos -como John Searle- la capacidad de pensamiento solo puede darse en un sistema que posea los mismos poderes causales que el cerebro (Searle, 1983:468) (lo cual es, o bien una indicación demasiado trivial, o

bien demasiado vaga); para otros -como David Chalmers- la conciencia que acompaña al pensamiento es una propiedad concomitante de los procesos funcionales de ciertos sistemas muy especializados, aunque virtualmente irreductible a esos puros procesos funcionales (Chalmers, 1999). Estas definiciones no permiten visualizar si una máquina digital *artificial*[2] podría contener esos poderes causales á la Searle o si podría verse acompañada de esa propiedad concomitante e irreductible á la Chalmers. ¿Mediará una cuestión de tiempo, investigación y tecnología para crear esas propiedades / poderes causales, o más bien estas definiciones nos cierran dicha posibilidad? Como señala Dennett acerca de los "misteriosos poderes causales del cerebro" que invoca Searle, es posible que otros seres humanos no los tengan, o tengan principios muy diferentes (Dennett,1987:234). ¿Será esto una prueba de que podría haber sistemas de complejidad suficiente como para tener un pensamiento, pero que no piensan? ¿Qué pasaría si descubriésemos que los polacos tienen poderes causales muy diferentes a los del resto del mundo? Incluso, como bien afirma Putnam, un sistema no tiene por qué conocer los propios poderes causales que lo inducen a estar en un determinado estado, o a pasar de un estado a otro, a menos que otro sistema -sea externo, sea un subsistema- pueda escanear esos procesos y comunicarlos al sistema principal. Pero desde una perspectiva puramente de primera persona (puramente cartesiana), la posibilidad de afirmar de manera tajante que la estructura biológica humana es condición causal suficiente para crear consciencia es, por lo menos, problemática, pues dicha estructura donde se buscan los poderes causales solo puede obtenerse desde una perspectiva de tercera persona[3].

Chalmers parece tener un problema similar de perspectiva. ¿Cómo, si no desde una visión de primera persona, podemos corroborar -si es que podemos- que un sistema tiene esa extraña propiedad irreductible llamada *conciencia*? La conciencia no es deducible a partir del diseño, ni a partir de la complejidad, ni a partir de cualquier poder causal presente en los niveles más bajos. ¿De qué modo podríamos atribuir conciencia, entonces? Quizás soy el único ser consciente del universo, y la perspectiva en tercera persona no puede acudir en mi auxilio para zanjar esta cuestión: todo lo que veo allí son sistemas funcionalmente idénticos a mí -mejor dicho: eso es lo que pude inferir a partir de un escaneo *en tercera persona* que los subsistemas sensoriales han hecho de mi cuerpo, de otros cuerpos y de sus conductas-, pero no puedo deducir conciencia en ninguno de ellos[4].

Searle ha negado con entusiasmo y vigor que el test de Turing pueda darnos una explicación relevante del fenómeno de la conciencia. Chalmers sería más cauto en este punto, aunque tal vez podría invocar que cualquier computadora está acompañada de un aunque sea minúsculo estado de conciencia concomitante a cualquier proceso de información.

¿Y dónde está el modelo?

El desafío de Turing puede verse desde otra perspectiva que podría evitar en parte los problemas semánticos y metodológicos de su planteo inicial. A la pregunta "¿qué sucede cuando una máquina juega al juego de la imitación?" podemos trocarla por otra que pueda captar una clase de contraste más interesante y que, de paso, ponga en juego no solo los poderes verbales de una máquina digital, sino también la capacidad de encajar con nuestras teorías acerca de cómo funciona la mente. En otras palabras: modificaremos el test para que el humano al otro lado del teletipo pueda ver cómo fue diseñada la máquina, tanto desde el punto de vista del hardware como del software. Esto es cambiar el juego completamente, y requerirá no solo de la competencia lingüística promedio del "participante", sino también de una enorme destreza técnica: el participante deberá ser un *experto*. En otras palabras: no nos conformaremos con el sospechoso "engaño" que la máquina puede hacer o no; queremos ir más allá de dicho engaño y desentrañar los mecanismos funcionales de la computadora.

Para ello debemos preguntar si la máquina en cuestión es un *modelo* de la mente. Reemplazamos, entonces, la pregunta de Turing por esta otra: "¿Encontrará el experto que la máquina es un modelo (no trivial) de la mente?"

Esta pregunta ha sido contestada de manera enérgicamente negativa por parte de quienes afirman que *necesariamente* una máquina artificial nunca puede tener una mente. Sin embargo, para que algo sea un modelo no debe ser idéntico a lo modelado. Además, la noción de modelo no presupone características físicas del sistema, sino características estructurales: el modelo empírico requiere de un modelo formal. Los pasos, entonces, para buscar un modelo de las cosas que nos rodean en el mundo son los siguientes: partimos de un sistema dado; estudiamos su historia[5], elaboramos su estructura -cuanto más detallada sea la historia, más precisa será la estructura- o modelo formal; a partir de allí elaboramos una teoría y, finalmente,

elaboramos un modelo empírico. A partir de este modelo, podemos diseñar -hipotética o actualmente- otros sistemas que cumplan con la estructura de dicho modelo. Un modelo -sea formal o empírico- puede ser instanciado (parcial o totalmente) por una multitud de sistemas en la naturaleza, o por uno solo o por ninguno. Dado que existiría la posibilidad de que no se encontraran sistemas naturales que cumplan con el modelo, salvo quizás en un único caso difícil -por ejemplo, el del cerebro humano-, podemos elaborar sistemas menos complejos -aunque con crecientes grados de complejidad- que sirvan de modelo a los variados procesos funcionales del cerebro humano. Si disponemos de una teoría -buena o mala- acerca del funcionamiento de los procesos cognitivos, podríamos elaborar un modelo para instanciar dicha teoría en un diseño artificial. Exactamente esto es lo que ha hecho Colby con su programa paranoide Parry: se basó en teorías acerca de la paranoia, especialmente en el enfoque de procesamiento de información de S. S. Tomkins (1963). Tomkins sugiere que el paranoico humano está en estado de vigilancia permanente intentando maximizar la detección de insultos y de minimizar la humillación: he aquí el componente teórico. El programa paranoide rastrea las oraciones en busca de pistas que sugieran daños explícitos e implícitos. Algo similar puede decirse del programa ELIZA de Weizembaum, el cual pretende imitar la estructura de diálogo de una psicóloga no directiva cuya descripción teórica clásica -repetimos: buena o mala- se encuentra en Rogers (1951). Bastan unas pocas líneas de diálogo con PARRY o con ELIZA para que salte a la vista la pobreza conceptual y la imposibilidad manifiesta de atribuirle conciencia a *esos* programas. Sin embargo, eso no refleja que el modelo proyectado deba ser necesariamente incorrecto (desde luego, lo es si los basamentos teóricos en los que se apoya dicho modelo son manifiestamente erróneos); podemos elaborar sistemas muy simples que solo den cuenta de determinada estructura funcional en algunos aspectos muy reducidos, dejando de lado otras cuestiones más complejas: esta es una estrategia útil para aislar hipótesis. Por ejemplo, PARRY pretende ser un paranoico puro[6], con lo cual, si tanto la teoría como la implementación son correctas, se podrían deducir cuestiones importantes relacionadas con la paranoia en general.

En la pregunta formulada más arriba, "¿Encontrará el experto que la máquina es un modelo (no trivial) de la mente?" pusimos entre paréntesis la "no trivialidad". Esta es una restricción cautelosa, pues los modelos triviales de la mente se obtienen a partir de descripciones muy generales y corremos el riesgo de que, si partimos de modelos tan bastos, encontremos que ciertos *agregados* -que ni siquiera entran en la categoría de *sistemas*- se "comporten" de acuerdo a ese modelo (quizás puede haber un nivel de descripción en el cual el cerebro y los papeles amontonados por el viento tengan algo en común). Pero esta restricción pone en riesgo toda la argumentación anterior: ¿Cómo distinguir los modelos relevantes -epistémicamente relevantes- de aquellos que no lo son? Aquí se debe tomar una decisión metodológica para lo cual se deberá tener en cuenta qué tan exhaustiva es la historia del sistema modelado, y qué tan detallada sea la teoría elaborada a partir de esa historia. Daniel Dennett(1987)ofrece un interesante y muy ilustrativo método de decisióndebemos proyectar diversos tipos de actitudes hacia los sistemas, y estas actitudes nos devolverán el tipo de descripción que necesitamos. Existen tres tipos de actitudes básicas: la actitud *física*, según la cual solo nos contentamos con la historia física de un sistema; la actitud de *diseño*, según la cual nos interesan sus componentes funcionales y la actitud *intencional*, desde la cual esperamos encontrar un sistema que posee creencias y actitudes proposicionales. Un grupo de papeles amontonados por el viento puede describirse, sin pérdida de información -esto es: sin dejar de lado detalles importantes de la *historia*[7] - desde la actitud física. Los programas PARRY y ELIZA, si bien parecen ser intencionales en una primera ojeada, nos muestran a través de su frecuentación que solo utilizan un número limitado de rutinas que pueden ser descriptas sin pérdida[8] por la actitud de diseño. Finalmente, los seres humanos y los animales se comportan como si tuvieran creencias y en esos casos les atribuimos la actitud intencional.

Despejar el camino: proyectos y conclusiones

Un modelo no trivial de la mente no tiene por qué ser un modelo *total* de la mente y un modelo de la mente -sea total o sea parcial- no tiene por qué pensar. La modelización de las características funcionales relevantes de algo no tiene por qué reproducir los efectos de dichas características. Un modelo es puramente virtual y sus efectos solo se pueden entender -parcialmente al menos- cuando se lo implementa en un sistema ¿Podemos esperar, entonces, que la *implementación* de un modelo muy detallado sí produzca el esperado chispazo? Esta es una de las preguntas clave de la filosofía de la mente. Es probable que Turing diga "sí" y Searle diga "no". Sin embargo, ambos autores -famosos por sus experimentos mentalesomitieron algo en sus experimentos que podía ayudar a encender el fuego: el contacto de la "máquina" con el mundo exterior. El hombre encerrado en el cuarto chino de Searle, al igual que la computadora de Turing, no tienen medio de saber qué pasa afuera. La máquina de Turing, porque no posee sentidos. El "demonio de Searle", porque está confinado a una habitación. Al igual que PARRY, el hombre encerrado en la habitación china no

comprende los símbolos de entrada; solo los manipula de acuerdo a reglas establecidas. Esto lo expresa Searle diciendo que de la sola sintaxis no puede esperarse comprensión; esto es: semántica. Ahora bien, la máquina de Turing podría remedar su incapacidad para responder acerca de cuestiones básicas como "¿Sientes que hace calor aquí?" con el agregado de métodos de escaneo y pautas sensoriales (dejemos de lado el problema de la viabilidad empírica de tales métodos). Si agregamos un método de escaneo, le hemos dado al sistema algún tipo de "semántica natural": el sistema debe interpretar la estimulación fotónica, o la vibración de las moléculas, o lo que fuere para lo cual el método de escaneo esté preparado. Cuanto más rico, detallado y ágilmente actualizable sea el escaneo, y cuanto más rápido y mejor pueda procesar dichos detalles, el sistema podrá ser más fiable en sus respuestas, las cuales deben surgir sobre la base de una interpretación inicial. Podría replicarse que esta no es una auténtica semántica. Dennett, sin embargo, postula que no existe tal "semántica de verdad"; todo lo que tenemos son estructuras sintácticas que *se comportan como* si produjeran estructuras semánticas[9]. El procesamiento de una multitud de estímulos por parte del cerebro y su correspondiente "contaminación" con el lenguaje forman un sistema complejo que crea (inventa, proyecta) una semántica ulterior a partir de esa semántica natural de los estímulos. Según Dennett, somos sistemas satisfactores, no optimizadores. Esto significa que nuestro cerebro tiende a completar información faltante y a sacar conclusiones apresuradas porque, como productos que han evolucionado a partir de un contacto rico, directo y continuo con su ambiente, debemos tomar decisiones rápidas aunque no sean lógicamente perfectas. Esa es, entonces *toda* la semántica que necesitamos postular: la que extraemos a partir de los estímulos del medio ambiente, dentro del cual está incluido el medio social. Un modelo funcional de la mente no puede dejar afuera el contacto con el medio. Por ello, si queremos responder a nuestra pregunta de manera completa, "¿Encontrará el experto que la máquina es un modelo *total* (no trivial) de la mente?", primero debemos asegurarnos de que esa máquina tenga un contacto tan rico con el medio ambiente como lo tenemos todos aquellos a los que nos dirigimos adoptando la actitud intencional. Según StevanHarnard(1989) para poder dar fiabilidad a este test, debemos elaborar un "test deTuring total", en el cual no solo se pondrían en consideración las capacidades computacionales -la capacidad de responder a preguntas que requieren cierto desarrollo deductivo o inductivo- sino también las *robóticas* -las que surgen del escaneo del ambiente-.

Lo segundo que debemos hacer es cerciorarnos de que la teoría utilizada no contiene errores, que es lo suficientemente detallada y exhaustiva para no resultar trivial, que el modelo empírico elaborado es correcto y que la implementación se ajusta a todo el proceso anterior. Para ninguno de estos pasos hay un camino despejado, pero gran parte de los estudios filosóficos de los últimos cincuenta años apunta a desmalezar la arquitectura conceptual folk que se interpone entre la descripción de los sistemas intencionales y la subsecuente teoría.

Hemos modificado el experimento de Turing por otro; y consiguientemente reemplazamos la pregunta. No nos interesa qué opina un sujeto de prueba particular y no calificado; nos interesan las opiniones de un experto. Tampoco abordamos la pregunta acerca de si la máquina piensa, pero sí podemos establecer si se trata de un modelo no trivial del pensamiento de acuerdo con las mejores teorías vigentes acerca de lo mental.

Apéndice: Wittgenstein y la proyección de la actitud humana

Mientras una buena parte de los esfuerzos actuales ahonda en el desmantelamiento de los conceptos de psicología popular, otra corriente nacida a partir del Wittgenstein de las *Investigaciones Filosóficas* tiende a reivindicar tales conceptos. Para Wittgenstein, según la interpretación que hace Kripke(2006), las personas tienen una mente, poseen pensamientos, creencias, temores y deseos porque se los atribuimos por medio de una actitud por la cual decidimos interpretar al sujeto de ciertas conductas como *humano*. Esa actitud es *preproposicional* y no surge a partir de una evaluación de las conductas; tiene un carácter espontáneo y consiste básicamente en una proyección más que en una inducción a partir de casos.

Si a los tres tipos de actitud que enunciaba Dennett (física, de diseño, intencional) le agregamos este cuarto tipo, el de la actitud prelingüística de *atribución de humanidad*, no solo encontramos seres intencionales sino que además entre ellos a algunos los calificaremos de personas. Este último tipo de actitud no contemplada por Dennett puede convertirse en crucial para distinguir cualquier ser intencional de uno que, específicamente, tiene objetivos humanos y que reclama para sí un reconocimiento diferente con respecto a otros seres intencionales.

Notas

[1] Podría pensarse, desde luego, que Turing no fue partidario del verificacionismo. Pero, en ese caso, ¿por qué habría de calificar de "sinsentido" a la noción de pensar, sin más análisis? Su método tiene todo el aspecto del método carnapiano, según el cual

aquello que no puede reducirse a un lenguaje protocolar es un sinsentido. Es posible, también, que Turing solo se haya comportado como verificacionista en *este* punto y que, por lo demás, no compartiera tal concepción. Si es así, solo resta señalar que, desde fuera del verificacionismo, el término pensar posee pleno significado y que, fuera de todo sesgo académico, existen seres a quienes tradicionalmente se les atribuye pensamiento y que, para que se les pueda atribuir tal predicado, es necesario que cumplan fuertes requisitos objetivos. Esos seres son, desde luego, las personas y probablemente algunos animales, los cuales son, precisamente, aquellos a quienes Turing no les atribuye el predicado "ser una máquina".

[2] Aquí recalcamos el término "artificial", dado que tanto Searle como Chalmers admiten que el cerebro es una máquina digital de procesos masivamente paralelos. En ambos casos, admiten dentro de la definición de "máquina" aquello que Turing dejó explícitamente fuera.

[3] A la perspectiva de tercera persona -en contraposición de la "fenomenología" o "perspectiva de los qualia"-Dennett la denomina "heterofenomenología".

[4] Chalmers no analiza el solipsismo como un caso plausible y relevante. La plausibilidad, según Chalmers, está dada por la "no arbitrariedad" de alguna propiedad. ¿Por qué -argumenta- existirían muchos sistemas funcionalmente semejantes a mí mismo, pero yo habría de ser el único de esos sistemas que posee conciencia? La línea que divide a "mí mismo" de los otros debe trazarse, a su juicio, de manera arbitraria y requiere una ulterior argumentación. Sin embargo, admite que la plausibilidad solo es parte de una presuposición general de *coherencia* e *invariancia organizacional* que no puede probarse empíricamente y que, por lo tanto, debe aceptarse como un supuesto epistémico. Si no se presupone una cosa tal como la coherencia entre la conciencia y la psicología, podríamos encontrar que el mundo se comporta de manera arbitraria, y de este modo jamás podríamos especificar leyes. Los presupuestos epistémicos funcionan como una *apuesta* de que la realidad funciona mediante leyes y no de manera arbitraria (Cf. Chalmers, 1999, tercera parte, cap. 6 y 7).

[5] Historia" entendida como la descripción de las propiedades de un sistema.

[6] Otra cuestión es, por supuesto, preguntar si tal modelo fue correctamente implementado o no en el diseño.

[7] La palabra "historia" se utiliza solo para hablar de sistemas, pero un grupo de papeles no entra en esta categoría. Por lo tanto, la estoy usando aquí en un sentido marginal, como la descripción de cada uno de los componentes de ese agregado contingente y no modelable.

[8] Nuevamente podríamos replicar: "¿Qué quiere decir aquí "sin pérdida"? ¿Cómo puedo saber si no estoy dejando de lado alguna descripción del sistema que es relevante?" Una vez más, aquí debemos tomar una decisión metodológica y argüir que sabemos inductivamente qué esperamos encontrar en un sistema dado. No es la mejor estrategia -después de todo, podría ocurrir que hasta los sillones tuvieran pensamientos profundos, como parodiaba Wittgenstein, pero resulta una buena apuesta a partir de nuestro contacto con el mundo. Quizás PARRY tenga intenciones y deseos reales (sea lo que se quiera decir con "real" y "tenga"), pero dado que estas actitudes son jerárquicas -un sistema intencional está "montado" sobre un sistema previamente "diseñado"lo único que hacemos es ser cautelosos si nos detenemos en la historia del diseño y dejamos de lado (por dudosa, o por lo que fuere) la supuesta historia intencional.

[9] En *La Actitud Intencional*, Dennett no es demasiado explícito con esta aseveración y solo se limita explicarla mediante analogías y ejemplos.

Bibliografía

1. Boden, Margritte (1984), *Inteligencia artificial y hombre natural*, Madrid, Tecnos, p. 143.
2. Chalmers, David (1999), *La Mente Consciente*, Barcelona, Gedisa
3. Dennett, Daniel (1987), *La Actitud Intencional*, Barcelona, Gedisa, p. 294.
4. Harnard, Stevan (1989), "Minds, Machines and Searle" en *Journal of Theoretical and Experimental Artificial Intelligence*, 1, pp. 5-25.
5. Kripke, Saul (2006), *A propósito de reglas y lenguaje privado*, Madrid, Tecnos.
6. Putnam, Hilary (1984), "Mentes y Máquinas", en Alan Ross Anderson (edit.), *Controversia sobre Mentes y Máquinas*, Barcelona, Tusquets, p. 113.
7. Rogers, Carl (1951), *Client Centered Therapy: Current Practice, Implications and Theory.* Boston, Houghton Mifflin.
8. Searle, John (1983), "Mentes, cerebros y programas", en Douglas Hoffstadter y Daniel Dennett,*El Ojo de la Mente,* Sudamericana, Buenos Aires, p. 468.
9. Tomkins, Silvan (1963), *Affect, Imaginery, Consciousness, Vol II: TheNegativeAffects*, Nueva York, en Boden, Margritte (1984), *Inteligencia artificial y hombre natural*, Madrid, Tecnos, p. 136.
10. Turing, Alan (1984), "Maquinaria, Computadora e Inteligencia", en Alan Ross Anderson (edit.), *Controversia sobre Mentes y Máquinas*, Barcelona, Tusquets, p. 11.
11. Van Fraasen, Bas (1977), "The pragmatics of explanation", en Clarendon P., *American Philosophical Quarterly*, Oxford, 14, p. 1143 - 1150.
12. Ziff, Paul (1960), *Semantic Analysis*, Cornell University Press, Nueva York, en Putnam, Hilary, *Mentes y Máquinas*, en *Controversia sobre Mentes y Máquinas* (1984), Alan Ross Anderson (edit.), Barcelona, Tusquets, p. 113.

Universidad Nacional del Sur

Departamento de Humanidades
Universidad Nacional del Sur
12 de Octubre y San Juan, 5to. piso
8000 - Bahía Blanca - Buenos Aires
República Argentina
Tel.: (54 291) 4595114

cdelsur_filosofia@uns.edu.ar