ISSN: 1576-3935

De la Habitación China al Laboratorio de I.A.

Mario SANTOS SOUSA

Universidad Autónoma de Madrid

Recibido: 26/09/05
Aprobado: 3/11/05

¿Es legítimo atribuir "mente" a un interlocutor mecánico (demostrablemente) competente? El experimento mental de la 'Habitación China', concebido por Searle en 1980, ha generado un torrente de literatura, de réplicas y contrarréplicas, que llega hasta nuestros días. Veinticinco años después, ¿queda algo interesante que añadir a la discusión, aparte de resucitar viejas objeciones? ¿O acaso hemos alcanzado un impasse filosófico, una suerte de equilibrio insulso? A mi modo de ver, la posición de Searle plantea, aún hoy, un serio desafío al sueño de una Inteligencia Artificial fuerte. No obstante, considero que dicho desafío no es en modo alguno insuperable. Para ello, propongo adoptar un punto de vista diferente, no el de la Habitación China sino el de un laboratorio de I.A.

Imaginemos que hubiese «máquinas tales que semejasen a nuestros cuerpos e imitasen nuestras acciones» (Descartes [1637] 2001: 88). ¿De qué manera podríamos reconocer que no por ello son *humanas*? El propio Descartes ofrece una posible respuesta:

> «Pues si bien se puede concebir que una máquina esté de tal modo hecha que profiera palabras, y hasta que las profiera a propósito de acciones corporales que causen alguna alteración en sus órganos, como, *verbi gratia*, si se la toca en una parte, que pregunte lo que se quiere decirle, y si en otra, que grite que se le hace daño, y otras cosas por el mismo estilo, sin embargo, no se concibe que ordene en varios modos las palabras para *contestar al sentido de todo lo que en su presencia se diga*, como pueden hacerlo aun los más estúpidos de entre los hombres» (Descartes [1637] 2001: 88; la cursiva es mía).

Descartes señala que la principal característica que nos distingue como humanos, es decir, nuestra inteligencia, no depende exclusivamente de nuestras acciones corporales, sino de nuestra capacidad para mantener una conversación en la que el sentido de nuestras palabras se ajuste al sentido de las palabras proferidas por nuestro interlocutor. El 'juego de imitación' que propone Turing en "Computing Machinery and Intelligence" (1950) recoge esa misma idea: se trata de un test que pone a prueba la competencia lingüística de un interlocutor mecánico. También conocido como 'Test de Turing', consiste en que una máquina convenientemente programada mantenga una conversación con un experimentador (quien debe averiguar si su interlocutor es humano o mecánico, *solamente* en función de cómo responde al sentido de sus preguntas). Aunque Descartes no pudiese concebir un agente mecánico capaz de superar semejante test, es decir, de "engañar" al experimentador haciéndose pasar por "humano", ¿qué razones tenemos para descartar dicha posibilidad?

Merece la pena examinar la postura que defiende Searle frente a la llamada Inteligencia Artificial fuerte, ya que pone en entredicho esa misma posibilidad (tomada en su sentido *fuerte*). Ante la idea de que una máquina adecuadamente programada no sólo pueda *simular* la inteligencia humana, esto es, mantener una conversación inteligente, sino *realmente comprender* lo que hace y dice, además de poseer otros estados mentales (o intencionales), Searle objetará que comprender un lenguaje (o tener estados mentales en general), no sólo consiste en realizar una tarea de tipo simbólico, sino que «requiere tener una interpretación, o un significado unido a dichos símbolos» (Searle 1984: 33): interpretación de la que simplemente carecería una máquina.

Searle ilustra su crítica mediante el experimento mental de la 'Habitación China', que presenta por primera vez en su artículo "Minds, Brains, and Programs" de 1980: supongamos que nos encierran en una habitación y nos entregan un manual que contiene reglas para manipular ideogramas chinos (o, mejor aún, símbolos marcianos). Acto seguido, nos hacen llegar del exterior secuencias de símbolos que debemos transformar (utilizando el manual) para formar nuevas secuencias que, finalmente, enviaremos de nuevo al exterior. Desde *fuera* parece que estemos manteniendo una conversación inteligente en chino (¡o en marciano!), mientras que *nosotros* sencillamente ignoramos que esos *in-* y *outputs* reciben el nombre de 'preguntas' y 'respuestas' respectivamente... ¡ni siquiera sabemos que estamos manteniendo una conversación (menos aún sobre qué trata)! Análogamente, un interlocutor mecánico recibe un *input*, lo procesa o computa siguiendo una serie de reglas, y completa su tarea mecánica e inconsciente emitiendo un determinado *output*. Al igual que «la persona encerrada en la habitación, dispone de toda la sintaxis que le podamos dar, pero no por ello adquiere la semántica apropiada» (Searle; en Dennett 1987: 323).

La Habitación China es un experimento mental que apela a la intuición del lector, invitándole a extraer la conclusión de que "tener un programa –un programa por sí solo– no es suficiente para (ni equivalente a) tener una mente" (Searle 1984: 39). Sin embargo, merece la pena estudiar las premisas que subyacen al *Gedankenexperiment* de Searle:

1. Un programa se define enteramente por su estructura formal o sintáctica.
2. Una mente tiene contenidos mentales o semánticos.
3. Tener sintaxis no es suficiente para (ni equivalente a) tener semántica.

De acuerdo con Searle, la primera es verdadera por definición, «es lo que entendemos por programa». Mientras que la segunda «refleja un hecho evidente del funcionamiento de nuestras mentes»: los estados mentales son intencionales, versan *sobre* (o refieren *a*) algo, por ejemplo, ciertos estados de cosas en el mundo. Finalmente, la tercera es una verdad conceptual que «simplemente articula nuestra distinción entre la noción de lo que es puramente formal y lo que posee un contenido» (Searle 1984: 39).

No obstante, dichas premisas requieren ser examinadas con mayor atención. En primer lugar, cabe preguntarse si es correcto que un programa se defina "enteramente" por su estructura formal. Dennett señala adecuadamente que «si no se fijan algunos detalles de implementación –por medio de la semántica interna del lenguaje de máquina en el que el programa está escrito en último término– un programa ni siquiera es un objeto sintáctico, sino meramente un conjunto de marcas tan inerte como un papel estampado» (Dennett 1987: 337). En efecto, parte de la "identidad" del programa viene dada por lo que el programa *hace*, por cómo interactúa con el mundo y lo modifica. (Pensemos en un virus informático, por ejemplo. Su capacidad destructiva constituye, sin duda, uno de sus rasgos definitorios.) Una vez examinada la tercera premisa, veremos en qué sentido la primera resulta inadecuada.

La segunda premisa descansa sobre el supuesto de que la mente humana posee estados intencionales "genuinos" (reales, intrínsecos, originales,...), mientras que un artefacto sólo posee una intencionalidad "derivada" (ilusoria o "*como-si*"). Sólo los seres humanos poseen mente, debido a las propiedades causales que tiene el cerebro humano. Searle es bastante explícito al respecto: «La intencionalidad es, en todo caso, un fenómeno biológico, y es tan probable que dependa de la bioquímica específica de sus orígenes como la lactancia, la fotosíntesis, o cualquier otro fenómeno biológico» (Searle 1980; en Hofstadter y Dennett [1981] 2000: 372).

Dos cuestiones ocupan nuestra agenda inmediata. Primero, ¿en qué medida depende nuestra mente de las "propiedades causales" del cerebro? Y segundo, ¿acaso se sostiene la distinción entre intencionalidad "genuina" e intencionalidad "derivada?"

Searle aboga por una suerte de "carbo-centrismo"[1] al declarar que es probable que la intencionalidad dependa causalmente de la "bioquímica específica" del cerebro. A continuación, trataré de destacar la importancia que puede llegar a tener la composición material del cerebro para nuestra inteligencia. Y ello en relación con algo que sí desempeña un papel esencial para la misma: el tiempo que tardamos en procesar la miríada de estímulos que nos llegan. De hecho, el cerebro humano alcanza velocidades de computación vertiginosas; dada su arquitectura en paralelo es capaz de realizar varias tareas al mismo tiempo: percibir, deliberar y actuar en función de un entorno complejo y en continuo cambio. De acuerdo con un estudio realizado por Sejnowski en 1985 (véase el artículo "Fast Thinking" de Dennett, 1987: 387), el cerebro tiene una capacidad media de procesamiento de 10^{15} operaciones por segundo, cinco órdenes de magnitud superior a los ordenadores digitales de procesamiento paralelo que existían en los ochenta. (Cabe notar, sin embargo, que un ordenador en serie podría realizar en principio las misma tareas... ¡a cambio de reducir considerablemente su velocidad!)

No obstante, todavía no hemos logrado demostrar por qué las *propiedades causales* del cerebro son relevantes. Aun suponiendo que algún día una máquina convenientemente programada alcanzase nuestra velocidad de procesamiento, ¿sería eso suficiente como para

1 Se puede encontrar la expresión, aunque empleada en otro contexto, en el artículo "Saving Machines From Themselves: The Ethics of Deep Self-Modification" de Peter Suber.

dotarla de *mente*? Searle es tajante: «No estamos hablando sobre un estado tecnológico particular. No tiene nada que ver con la distinción entre procesamiento serial o paralelo, o la velocidad de las operaciones de un ordenador, o con computadores que puedan interactuar causalmente con el entorno, o incluso con la invención de robots» (Searle 1984: 36-37). ¿Con qué otra cosa puede tener que ver? De acuerdo con Searle, tiene que ver con la distinción entre intencionalidad "genuina" e intencionalidad "derivada".

Incluso un interlocutor mecánico competente debe sus habilidades a un grupo de sesudos diseñadores humanos. Ningún programa de ordenador, ningún robot que podamos diseñar y construir, aunque veamos en él a un interlocutor inteligente, será jamás un pensador verdaderamente autónomo con el mismo tipo de intencionalidad genuina de la que gozamos los seres humanos.

Sin embargo, como trataré de mostrar a continuación, la distinción entre intencionalidad genuina e intencionalidad derivada difícilmente se sostiene. Pensemos, por ejemplo, en algún potente programa de ajedrez, como *Fritz* o *Deep Blue*. En cada jugada, el 'programa madre' debe evaluar un sinfín de movimientos. Para ello, ejecuta un gran número de subrutinas (y de sub-...-sub-rutinas...) que compiten entre sí por formar parte de la jugada ganadora. Cada una de ellas sigue su propio juego y diseña su propia estrategia, de tal manera que cada una posee una suerte de intencionalidad, aunque *derivada* con respecto al 'programa madre' que supervisa y evalúa (y, en cierta manera, sabe interpretar) sus movimientos para incorporarlos a *su* jugada. ¿Acaso podemos afirmar que el programa madre posea una intencionalidad *genuina*? Sí y no (sí con respecto a una determinada subrutina, no con respecto a su programador, generalmente humano): la *atribución* de uno u otro tipo de intencionalidad dependerá del punto de vista que adoptemos.

Ahora bien, demos un paso más y traslademos este ejemplo al caso que nos ocupa, adoptando un punto de vista estrictamente evolutivo. ¿En qué medida podemos afirmar que nosotros, los seres humanos, gocemos de un tipo de intencionalidad genuina? Cabe preguntarse si no somos también poseedores del tipo de intencionalidad que Searle consideraría "derivada", en tanto que somos producto (una subrutina, si se quiere) de un programa madre denominado 'Madre Naturaleza'. Pero si esto es así, ¿qué sentido tiene mantener una distinción *absoluta* entre intencionalidad "genuina" e intencionalidad "derivada"? Desde esta perspectiva, *todo* agente poseería una intencionalidad derivada.

¿Pero si aceptamos esto, no estamos cayendo en una suerte de pan-psiquismo? «El precio de negar la distinción entre intencionalidad intrínseca y derivada –dirá Searle– es el absurdo, porque convierte todo lo que hay en el universo en mental» (1992: 81). No obstante, negar dicha distinción *no* nos otorga licencia para atribuir mente a un termostato o a una piedra, en contra de lo que pueda pensar Searle. Para no caer en el absurdo, no debemos perder de vista el tema que nos ocupa, a saber, si es legítimo atribuir mente a un interlocutor mecánico (demostrablemente) competente (por ejemplo, capaz de superar el Test de Turing). Nadie está sugiriendo, pues, que sea posible mantener una conversación inteligente con un termostato o con una piedra.

Pasemos a la tercera premisa: tener sintaxis no es suficiente para (ni equivalente a) tener semántica. Ya hemos mencionado que un programa sin implementar es tan inerte como los estampados de un tapiz. De nuevo, es importante señalar que un programa se caracteriza en gran medida por lo que *hace*, por su capacidad de modificar estados de cosas. Este aspecto suscita cuestiones de interés legal. Imaginemos dos programas que cumplen la misma función, que realizan la misma tarea, pero que, sin embargo, difieren sintácticamente (p.ej., por estar escritos en lenguajes diferentes). ¿Se trata de un solo programa o de dos programas distintos? Ciertamente ambos "hacen lo mismo," aunque de distinta manera. Sea

como fuere, lo que queda claro es que resulta difícil separar el programa de sus condiciones de implementación. En cierto modo, declarar que un programa es incapaz de producir semántica por sí mismo es análogo a decir que una partitura no puede producir música por sí misma. Ambos tienen que ser ejecutados. Llegados a este punto, parece más adecuado preguntarse si un programa *implementado* es capaz de producir semántica.

Un ejemplo adecuado de un programa conversacional (implementado en una máquina suficientemente "ágil" como para interactuar a tiempo real con su entorno) es *Ripley*, el último "robot conversacional" del MIT Media Lab, un robot que tiene, entre otras capacidades, un sistema de percepción visual y auditiva, capacidades motoras, y la habilidad de manipular objetos de tamaño medio con su "boca mecánica." Además, Ripley es capaz de determinar la posición de un observador cercano, y utilizar dicha información para contextualizar la descripción de objetos en su entorno inmediato. Estas características le permiten realizar diferentes tareas, por ejemplo, comprender y ejecutar instrucciones del tipo 'Toca aquella cosa azul y pesada que se encontraba a mi izquierda', por tanto, es capaz de identificar objetos y manipularlos,... lo que pone de manifiesto un tipo de intencionalidad "cruda." Lo importante, no obstante, es que Ripley es capaz de asignar un equivalente funcional de "significado" a sus variables internas, interpretarlas y formar una representación interna de su entorno[2], en otras palabras (y sin pecar de optimismo): es capaz de producir semántica.

El experimento mental de la Habitación China constituye una "bomba de intuición" (*intuition pump*, tomando prestada la expresión de Dennett), destinada a estimular nuestra imaginación, y que nos invita a extraer la conclusión de que ningún programa por sí mismo es capaz de producir semántica, una conclusión que sitúa en un *impasse* a los defensores de la Inteligencia Artificial fuerte. No obstante, el tránsito de la Habitación China al laboratorio de I.A., arroja nueva luz sobre el problema, ofreciendo una perspectiva que, a mi modo de ver, resulta más fructífera e interesante, y que merece la pena explorar. Abre, en definitiva, la posibilidad de que una máquina convenientemente programada pueda producir lo que es capaz de producir un cerebro orgánico como el nuestro con sus poderes causales particulares: una mente competente capaz de desenvolverse *en* el mundo.

BIBLIOGRAFÍA:

DENNETT, D. C. (1987) *The Intentional Stance*. Cambridge: MIT Press/A Bradford Book.

DENNETT, D. C. [1985] (1998) *Brainchildren. Essays on Designing Minds*. Cambridge: MIT Press/A Bradford Book.

DESCARTES, R. [1637] 2001 *Discurso del Método*. (trad. Manuel García Morente). Madrid: Espasa, Col. Austral.

HOFSTADTER, D. R. y Dennett, D. C. [1981] (2000) *The Mind's I. Fantasies and Reflections on Self and Soul*. New York: Basic Books.

SEARLE, J. R. (1980) "Minds, Brains, and Programs" en Hoftstadter y Dennett [1981] 2000.

2 Para más detalles sobre el proyecto Ripley, consúltese la página http://www.media.mit.edu/cogmac/, en especial, el artículo introductorio "Conversational Robots: Building Blocks for Grounding Word Meanings" (Deb Roy et al. 2003).

SEARLE, J. R. (1984) *Minds, Brains and Science*. Cambridge: Harvard University Press.

SEARLE, J.R. (1992) *The Rediscovery of the Mind*. Cambridge: MIT Press/ A Bradford Book.

TURING, A.M. (1950) "Computing Machinery and Intelligence" en Hoftstadter y Dennett [1981] 2000.