Las objeciones de John Searle a la noción de inteligencia artificial y respuesta de D. Hofstadter

Luis A. Camacho Naranjo.
Escuela de Filosofía.
Universidad de Costa Rica

1. Cuando se analiza en qué consiste la inteligencia, dos son los enfoques globales más importantes: el biológico y el de la inteligencia artificial. El primero considera la inteligencia una propiedad material de un sistema físico caracterizado por determinado tipo de células, las neuronas. El segundo la considera una propiedad formal de cualquier sistema que reúna ciertos requisitos funcionales. Para el partidario del primer enfoque la pregunta misma "¿pueden pensar las máquinas?" resulta extraña e incluso irritante, pues parte de una asignación de propiedades a un sistema que casi por definición no puede tenerlas. Se comete entonces, según él, un error de categorías. Para el partidario del segundo enfoque, la pregunta no solo es plenamente legítima, sino casi ociosa: más bien habría que preguntarse cómo se aplica nuestro conocimiento de sistemas formales al funcionamiento del cerebro.

2. En 1950 Alan Turing publica en la revista **Mind** *(1)* su famoso artículo "Computing Machinery and Intelligence" en el que se plantea la pregunta "¿Pueden pensar las máquinas?" Turing encuentra que la pregunta está llena de implicaciones indeseables para la gente de su tiempo (aunque quizá no para otra gente en tiempos futuros), por lo que prefiere transformar la pregunta en otra diferente, a saber, la de si las máquinas pueden hacer lo que los seres humanos, caracterizados por su habilidad para resolver problemas complejos. Como es bien conocido, Turing desarrolla en este artículo su famoso "Turing test", que consiste en adivinar mediante preguntas el sexo de dos interrogados ocultos. ¿Qué ocurre –se pregunta Turing– si

en vez de un ser humano es una máquina la que contesta las preguntas? ¿Podrá engañar al interrogador del mismo modo que un interlocutor humano? Nos interesa señalar aquí que el mismo Turing anticipa los tipos de objeciones que se pueden hacer a su planteamiento y que uno de estos tipos (así llamados porque no son objeciones aisladas, sino esquemas de dificultades) tiene que ver -según Turing- con la conciencia, aunque sería más exacto decir que se refiere a las emociones.

El artículo de Turing marcó las pautas de la polémica sobre la existencia o no de inteligencia en las máquinas que llevan a cabo operaciones como calcular, jugar ajedrez, probar teoremas, etc. Treinta años después de su aparición, Douglas Hofstadter, profesor de ciencias de la computación en la Universidad de Indiana, publica su monumental obra **Goedel, Escher, Bach: An Eternal Golden Braid**,(2) que le valió el codiciado Premio Pulitzer. Convertida enseguida en un libro de gran venta, es una obra que innumerables personas mencionan y alaban aunque pocos parecen haberla leído en su totalidad. La enorme complejidad de la obra es una buena excusa. Hofstadter es un ferviente partidario de la noción de inteligencia artificial en el sentido más fuerte -pero también más sofisticado- del término, capaz de defender su posición mediante la construcción de un razonamiento por etapas en el que juegan parte importante no solo la lógica y la matemática sino también la anatomía y fisiología cerebrales y en el que se utilizan innumerables analogías tomadas de la música, la pintura y la literatura. Pienso que sin exagerar se puede decir que la noción de función recursiva es el tema que unifica toda la inmensa variedad de elementos de su argumentación. Gran admirador y seguidor de Turing, Hofstadter considera que el enfoque del creador del test para detectar inteligencia sigue siendo válido.

3. Al año siguiente de la publicación de **Goedel, Escher, Bach Hofstadter**, publicó una obra titulada **The Mind's I**,(3) que consiste en una colección de ensayos sobre el tema del yo de autores tan variados como Jorge Luis Borges, Stanislaw Lem, Thomas Nagel, etc. Muchos de estos ensayos son de extraordinaria calidad y gran originalidad. En particular nos interesa el de John Searle, profesor de lingüística en Berkeley, el cual destaca entre los demás ensayos por el reto que plantea a la noción de inteligencia artificial. El texto que recoge este volumen es el artículo titulado "Minds, Brains and Programs" aparecido en el número de setiembre de 1980 de **Behavioral and Brain Sciences**. (4) Dado que no parece ser conocido entre nosotros, y que la respuesta de Hofstadter que acompaña al trabajo de Searle merece un análisis, nos proponemos aquí dar a conocer tanto la argumentación del profesor de Berkeley como el comentario del profesor de Indiana. Luego añadiremos algunas reflexiones de nuestra propia cosecha.

Empieza Searle distinguiendo dos versiones de inteligencia artificial (en adelante usaremos IA) : la débil y la fuerte, digamos IA-D e IA-F. La IA-D considera a la computadora un instrumento útil para entender el funcionamiento del cerebro humano. Searle no tiene ninguna objeción que ha-

cer a esta versión -por lo menos en este artículo. La IA-F, en cambio, considera que las operaciones de la computadora en los programas más avanzados en nuestros días *son* operaciones inteligentes en el mismo sentido en que lo son las operaciones humanas equivalentes. Esta posición la resume Searle en la página 363 con la expresión: "la mente es al cerebro lo que el programa es a la máquina". Si las máquinas son capaces de responder preguntas sobre un tema, entonces literalmente *entienden* lo que se les ha dicho y esta actividad de las máquinas *explica* lo que hace el cerebro humano. Son estas afirmaciones las que Searle encuentra injustificadas, y lo que tiene en mente es el trabajo de Roger Schank y sus colaboradores de la Universidad de Yale (1977), así como el programa SHRDLU de Terry Winograd (1973), el ELIZA de Weizenbaum (1965) y, en general, cualquier simulación de fenómenos mentales humanos. El trabajo de Schank consistió en crear un programa que "entiende" relatos y es capaz de contestar preguntas relacionadas. Tomemos dos relatos: (a) "Un hombre entró en un restaurante y pidió una hamburguesa. Cuando se la trajeron estaba quemada. El hombre se levantó indignado y se fue sin pagar". (b) "Un hombre entró en un restaurante y pidió una hamburguesa. Cuando se la trajeron sonrió y pagó la cuenta que le dieron". Si se hace ahora la pregunta "¿se comió el hombre la hamburguesa?" La respuesta correcta sería "sí" en el primer caso y "no" en el segundo. Ahora bien: el programa de Schank hace posible que una máquina pueda responder estas preguntas correctamente a partir del relato correspondiente.

A diferencia de Hubert Dreyfus -quien en la década de los setenta atacó la idea de la inteligencia artificial desde la perspectiva de la fenomenología y acabó desprestigiado cuando una computadora le ganó una partida de ajedrez después de afirmar él que esto no era posible (5)-Searle utiliza el mismo marco de referencia del partidario de la inteligencia artificial. Es esto lo que hace su argumentación tan interesante.

Se establece ante todo un principio general: *"Una manera de probar cualquier teoría de la mente consiste en preguntarse qué ocurriría si la mente de uno funcionase según los principios que, de acuerdo con la teoría, rigen todas las mentes"*. (pág. 355). Apliquemos a continuación este test al programa de Schank y afines, mediante un experimento mental: supongamos que me colocan dentro de una habitación cerrada y me dan un largo texto escrito en chino, idioma que desconozco en absoluto hasta el extremo de que ni siquiera estoy seguro de poder distinguir entre algo escrito en chino o en japonés, y menos aun de distinguir entre un texto chino y una colección de garabatos sin sentido parecidos al chino. A continuación, me dan otro largo texto en chino y, además, un conjunto de reglas para correlacionar el primer texto con el segundo. Estas reglas están escritas en mi lengua nativa, cualquiera que ésta sea. Las reglas establecen correlaciones entre los signos del primer texto y los del segundo con base únicamente en la forma de los signos. Un paso más: me dan otra serie de signos en chino, y otra serie de instrucciones en mi lengua nativa que tiene que ver con la correlación del tercer texto chino con los dos anteriores. Estas instrucciones me permiten

devolver ciertos signos reconocibles por su forma, tomados de los dos primeros textos, en respuesta a ciertos signos que aparecen en el tercer texto. No tengo la menor idea de lo que significan los signos del tercer texto, ni los de los otros dos, ni por qué a los signos del tercero corresponden otros de los dos textos anteriores. Pero sí puedo seguir las reglas dadas en mi lengua nativa y estas reglas únicamente tienen que ver con manipulación de signos. Fuera de la habitación dentro de la que estoy, alguien -cuya identidad desconozco- llama a los dos textos chinos un "relato", a las instrucciones para relacionar los textos "un programa" y al tercer texto en chino "una serie de preguntas". Además, a los caracteres chinos que le devuelvo, siguiendo las instrucciones que recibo en mi lengua, las llama "respuestas". Después de un rato de ejercicios no me equivoco en seguir las instrucciones, es decir, no cometo errores en la correlación entre signos del tercer texto chino y los dos primeros. Supongamos que quien escribió las instrucciones, y los tres textos, lo hizo sin equivocarse. Ahora bien: para un observador fuera del sistema, que conozca chino, llegará un punto en que no podrá distinguir entre mis "respuestas" y las que daría un chino nativo, es decir, un hablante de la lengua desconocida para mí en la que están escritos los textos. Lo único que he hecho es seguir instrucciones, sin entender de qué se trata, pero para alguien fuera del sistema he entendido un relato en chino y he sido capaz de contestar correctamente preguntas sobre ese relato. Si no se explica lo ocurrido a alguien que examina mis "respuestas" en chino, concluirá lógicamente que conozco chino. Sin embargo, no es éste el caso, y la razón es que lo que he hecho es manipular signos siguiendo instrucciones sin conocer el significado de los signos. La manipulación es coextensiva con las instrucciones, y fuera de éstas no parece haber posibilidad de "acertar" en las "respuestas" a las preguntas formuladas.

Lo interesante de este experimento mental es que nos da una imagen clara de la diferencia entre entender y operar sin entender, aunque no sepamos aún en qué consiste entender. Si bien estamos muy lejos de saber cómo opera el cerebro, la argumentación de Searle se basa en una clara distinción entre la operación mental de *entender* algo y la de manipular signos correctamente, pero sin entender. Searle argumenta así: *Si nuestro cerebro funciona como lo postulan los partidarios de IA-F, entonces no entendemos. Pero es así que entendemos; luego los partidarios de IA-F están equivocados.*

4. Al igual que Turing, Searle recoge varios esquemas de objeciones y les da respuesta. Vamos a resumirlos, y para mayor facilidad les pondremos un breve epígrafe en vez de un nombre:

(a) "Hay que ver el sistema": el individuo en cuestión no entiende lo que hace, pero no es más que una parte de un sistema del cual sí se puede decir que entiende. Lo que interesa es analizar todo el conjunto, y no una de sus partes. Respuesta de Searle: nada cambiamos con hacerlo, puesto que el individuo puede interiorizar todos los elementos del sistema aprendiéndose de memoria todas las instrucciones e incluso los signos en chino, y hacer luego las operaciones en su mente. Esto no hace variar en lo más mínimo el hecho de que manipula signos de cuyo significado no tiene noción.

(b) "Pongamos una computadora dentro de un robot": ahora la computadora controla el robot, y éste tiene percepción de la realidad exterior, e incluso puede moverse. El robot hace las operaciones, siguiendo indicaciones de la máquina. Este robot tendrá entonces estados mentales relevantes. Respuesta de Searle: no, no los tiene, puesto que seguirá operando sin entender. Dotar el robot de estas capacidades en nada cambia el planteamiento original, y suponer que una computadora dentro del robot puede hacer lo que hace el cerebro dentro del cuerpo humano consiste en cometer una falacia de petición de principio.

(c) "Combinemos todos los elementos anteriores, y el resultado será una máquina que entiende": para Searle esto no sería suficiente para decir que la máquina entiende, pues la complejidad de la nueva situación no introduce ningún elemento nuevo que no haya sido analizado antes.

(d) "Copiemos operaciones cerebrales, en vez de dar instrucciones para operar con signos": ésta es sin duda la más interesante de las objeciones. Consiste en lo siguiente: se modifica el programa de computación diseñado por Schank, para que pueda simular lo que ocurre en el cerebro de un nativo chino que ciertamente entiende lo que lee en su idioma. La máquina que opera con este programa recibe textos en chino, los procesa con una serie de operaciones que copian las operaciones de un cerebro que entiende, y responde en chino. Ahora bien: si se dice del nativo que entiende, también se deberá decir lo mismo de una máquina que reproduce las mismas operaciones físicas del cerebro, es decir, que es capaz de reproducir fielmente lo que hacen las neuronas cuando un hablante nativo opera con los caracteres de su propio idioma. Ante esta objeción la estrategia de Searle es doble: primero, indicar que este camino es justamente el que la tesis de la IA-F quería evitar al decir que la inteligencia se da dondequiera que tengamos ciertas estructuras formales, de modo que no era necesario conocer el funcionamiento del cerebro para poder producir inteligencia. Más aún, como se recordará la tesis de IA-F quería precisamente ahorrarse este camino largo diciendo que la operación de la máquina era inteligente en sentido estricto y *capaz de explicar* la operación mental. La objeción que se analiza parte del supuesto contrario: es preciso conocer el funcionamiento del cerebro para reproducir la inteligencia. ¿En qué quedan entonces las pretensiones de la IA-F? En segundo lugar, Searle se imagina una situación en que se cumplan las condiciones establecidas por el objetante pero no se obtenga el resultado deseado: en vez de operar con signos, el ser humano que recibe los textos en chino sin saber ese idioma opera, siguiendo instrucciones, un sistema de tubos y válvulas de chorros de agua que imitan fielmente la secuencia de eventos que tiene lugar en las sinapsis de las neuronas del cerebro de un hablante nativo. Esto no hace que el individuo en cuestión entienda chino, ni menos aún que los tubos y válvulas lo hagan.

5. Hofstadter empieza negando el punto de partida de Searle, a saber, el que un ser humano pueda hacer lo que se dice en el experimento. Por breves que sean los relatos, y sencillas las instrucciones, una operación como

la descrita presupone toda la complejidad de la competencia lingüística. Más difícil resulta aceptar la respuesta de Searle a la objeción según la cual es el sistema complejo el que entiende: memorizar todas las instrucciones, y los textos cuyo significado se desconoce, le parece inverosímil a Hofstadter.

Es ésta la solución que adopta Hofstadter, la que en el artículo de Searle se denomina "respuesta sistémica": "entender" se dice, según esta posición, de las operaciones de un sistema enormemente complejo cuyos elementos apenas podemos vislumbrar en nuestros días, pero de los cuales sabemos que están organizados por niveles. Cuando entendemos un idioma somos capaces de "ver" a través de él, es decir, de usarlo sin darnos cuenta. Justamente cuanto mejor lo entendemos menos nos damos cuenta de que lo estamos usando; pero esto es posible porque los elementos de diversos niveles (señales auditivas o visuales, neuronas y sustancias químicas en el cerebro signos y símbolos, etc) se han integrado de tal manera que el conjunto resultante opera sin problemas. Searle ve el sistema del que se habla en la primera objeción como *dos* individuos separados: el que pregunta y el que responde. Hofstadter lo analiza como un sistema individual visto desde dos puntos de vista, el de la operación de seguir instrucciones y el de la asignación de preguntas y recepción de respuestas. Así considerado, el sistema no entiende si lo vemos desde el primer punto de vista, aunque sí lo hace desde el segundo.

6. Hofstadter señala, con razón, que en ningún momento Searle nos ha indicado en qué consiste entender. Curiosamente, esta objeción se aplica también a quien la hace. Para quienes consideramos injustificadas las pretensiones de los partidarios de la *IA-F en el estado en que la investigación en este campo se encuentra en nuestros días* (6), la situación no podría ser de otro modo. Del experimento planteado por Searle concluimos que no se da inteligencia sin intencionalidad, y que por tanto no se puede hablar de inteligencia artificial en sentido estricto si no se ha resuelto el problema de la intencionalidad correspondiente. De la respuesta de Hofstadter, y del tema admirablemente desarrollado por él en **Goedel, Escher, Bach**, concluimos que la inteligencia presupone gran cantidad de elementos en niveles complejos cuyas propiedades no son reducibles a las de los niveles anteriores. La integración de los dos enfoques, el biológico y el formal, resulta entonces necesaria. Y a la distinción entre dos versiones de los partidarios de inteligencia artificial, hecha al comienzo de este artículo, debemos añadir otra: entre los partidarios ingenuos y los partidarios inteligentes de la inteligencia artificial. Asimismo, entre los críticos ingenuos y los críticos inteligentes. Hofstadter es uno de los proponentes más inteligentes; Searle uno de los críticos más perspicaces. Es este encuentro de habilidades lo que hace tan fascinante la discusión incluida en **The Mind's I**.

NOTAS

(1) Mind, vol. LIX, no. 236, 1950.

(2) Douglas R. Hofstadter Goedel, Escher, Bach. An Eternal Golden Braid. (New York: Basic Books, 1979; Vintage Books, 1980).

(3) Douglas R. Hofstadter–Daniel C. Dennett, eds. The Mind's I: Fantasies and Reflections on Self and soul. (Basic Books, 1981; Bantam Books, 1982). Las citas son de esta última edición.

(4) Vol. 3, Cambridge University Press, 1980.

(5) La principal obra de Hubert Dreyfus es What Computers Can't Do. (New York: Harper & Row, 2da. edición, 1979).

(6) Véase mi reseña titulada "Compulsive Technology: algunas dudas sobre la revolución de la computación y la así llamada "inteligencia artificial", en la revista Desarrollo (Costa Rica: ACOHIFICI–CITEPPOL), no. 5, agosto 1987, pp. 87–93.