Scientia*et*Fides 2(2)/2014 ISSN 2300–7648 / DOI: http://dx.doi.org/10.12775/SetF.2014.018

Received: September 5, 2014. Accepted: September 30, 2014

¿Basta la prueba de Turing para definir la "inteligencia artificial"?

(Is Turing Test enough to define artificial intelligence?)

MANUEL ALFONSECA
Profesor honorario de la Universidad Autónoma de Madrid
Manuel.Alfonseca@uam.es

Resumen. En los 64 años transcurridos desde que Alan Turing propuso su famosa prueba para definir la inteligencia artificial, ningún programa de ordenador se había aproximado a cumplirla. Ahora que este objetivo parece un poco más cercano y la prueba de Turing comienza a parecer insuficiente, conviene recordar los argumentos de John Searle en su contra.

Palabras clave: Inteligencia artificial; Prueba de Turing; Habitación china de Searle.

Abstract. Since 1950, when Alan Turing proposed his famous test to define artificial intelligence, no computer program has come near to fulfill it. Now that this goal seems a little closer and Turing test looks insufficient, it is convenient to remember John Searle's argumentation against Turing test.

Keywords: Artificial Intelligence; Turing Test; Searle's Chinese Room.

Alan Turing fue un científico polifacético, cuya investigación tocó muchos campos tan diferentes como la química, la biología, la informática y las matemáticas. En 1935 formuló el famoso *teorema de la parada*, que establece límites a la capacidad de cómputo y en el fondo equivale al teorema de Gödel, pero resulta más fácil de entender intuitivamente. También diseñó un tipo especial de máquina de cómputo que, en su honor, se llama *máquina de Turing*. Aunque no se construye con hardware, resulta fácil de simular con una computadora electrónica. Cada máquina de Turing está diseñada para ejecutar un solo programa. Sin embargo, Turing demostró que es posible diseñar una máquina especial (la *máquina de Turing universal*), que es capaz de simular el funcionamiento de cualquier otra máquina de Turing si se le proporcionan, como datos de entrada, la definición de esta y su *input* (los datos de entrada de la máquina simulada). La máquina de Turing universal tiene una capacidad de cálculo equivalente a la de nuestras computadoras electrónicas. En puridad, es aun más potente, porque se supone que tiene memoria infinita, pero esta diferencia se está haciendo cada vez menos importante, a la vista del tamaño que alcanza la memoria de las computadoras modernas.

Otra de sus más famosas contribuciones fue la llamada *prueba de Turing* para definir la inteligencia artificial, un criterio que propuso en 1950 para decidir si una máquina puede ser (o no) tan inteligente como el hombre. En una de las diversas formas en que se ha enunciado, la prueba de Turing (Santini 2012) viene a decir esencialmente que *si una máquina llegara a ser capaz de engañar a los seres humanos, haciéndose pasar por humana, con la misma facilidad con que un ser humano puede engañar a otro, habría que considerarla inteligente*.

El término *inteligencia artificial* fue inventado en 1956 por John McCarthy para dar nombre al campo de la informática que se dedica al estudio y al diseño de máquinas inteligentes. Algunos resultados preliminares prometedores en este campo (la demostración de teoremas matemáticos sencillos, así como programas capaces de jugar a juegos tradicionalmente considerados *inteligentes*, como las damas), llevaron a los investigadores de finales de los años cincuenta a lanzar las campanas al vuelo y a predecir que, en

solo diez años, sería posible construir programas capaces de ganar al campeón del mundo de ajedrez, y resolver de forma aceptable el problema de la traducción automática de un idioma a otro.

Pasaron los diez años y nada de eso ocurrió. El ajedrez resultó ser un juego mucho más complicado que las damas. En cuanto a la traducción automática, la ambigüedad sintáctica y semántica de las lenguas humanas, que además es distinta en cada una de ellas, hace casi imposible realizar una traducción perfecta sin disponer de una imagen global del mundo que las máquinas no poseen. Basta pensar en frases como estas:

- **Frases con ambigüedad sintáctica:** *Pasaré solo este verano aquí.* (*Solo* puede ser un adverbio o un adjetivo). *No pude estudiar derecho.* (*Derecho* puede ser un sustantivo o un adjetivo).
- **Frases con ambigüedad semántica (doble sentido):** *Nos vemos mañana en el banco.* (Puede referirse a un banco público en la calle o a una entidad bancaria).

El fracaso de las predicciones provocó el desánimo de la investigación en inteligencia artificial, pues muchos de sus practicantes se dedicaron a otras cosas. Sin embargo, aunque con altibajos, y no siempre con resultados satisfactorios, los avances continuaron llegando en un goteo continuo: sistemas expertos, redes neuronales artificiales, algoritmos genéticos... En 1997, 30 años después de lo previsto, un ordenador consiguió por fin vencer al campeón del mundo de ajedrez. (Dejamos aparte la cuestión de si el programa que lo consiguió puede realmente considerarse inteligente). También existen ya sistemas razonablemente buenos de traducción automática, aunque todavía es preciso que un ser humano repase sus resultados antes de utilizarlos, pues generan muchos errores. También ha avanzado mucho la conducción automática de vehículos (coches y aviones).

Entre tanto, la prueba de Turing sigue siendo, en general, inabordable para las máquinas. Cuando se dialoga con una de ellas, no se tarda mucho en descubrir que no estamos hablando con un ser humano, pues no consigue engañarnos. Este mismo año, por ejemplo, una universidad del Reino Unido organizó un concurso para programas que intentaran pasar la prueba de Turing, que fue ganado por un programa de respuesta automática de

chats (*chatbot*) llamado *Eugene Goostman*, que se hacía pasar por un chico ucraniano de 13 años y consiguió convencer al 33% de sus contertulios de que era un ser humano. El suceso ha tenido la consecuencia inmediata de que hayan surgido voces que sugieren que la prueba de Turing no es suficiente para definir la inteligencia, pues del análisis del programa se deduce (en opinión de esos críticos) que este es cualquier cosa, menos inteligente.

Como dice Evan Ackerman (Ackerman 2014), comentando el caso:

> El problema con la prueba de Turing es que realmente no demuestra si un programa de inteligencia artificial es capaz de pensar: más bien indica si un programa de IA puede engañar a un ser humano. Y los seres humanos somos realmente tontos. Caemos en toda clase de trampas que un programa bien hecho puede utilizar para convencernos de que estamos hablando con una persona capaz de pensar.

Es decir, que la prueba de Turing no basta para asegurar que una máquina es tan inteligente como nosotros. Hace falta algo más. En 1980, el filósofo John Searle propuso una nueva prueba: *la habitación china*. Veamos en qué consiste.

Supongamos que tenemos un programa de ordenador capaz de pasar satisfactoriamente la prueba de Turing dialogando (por ejemplo) con una mujer china. En la conversación, tanto la mujer como el ordenador se expresan en chino, utilizando caracteres chinos para comunicarse por escrito a través de un teletipo. El ordenador, que está encerrado en una habitación para que la mujer no lo vea, lo hace tan bien que es capaz de engañarla, por lo que la mujer creerá estar dialogando con un ser humano que conoce perfectamente la lengua china.

Searle propone lo siguiente: sacamos de la habitación al ordenador y en su lugar entra él (Searle), que no sabe chino, pero se lleva consigo un organigrama del programa que utilizaba la computadora para dialogar con la mujer. En principio, utilizando ese organigrama, Searle debería ser capaz de dialogar con ella en su propia lengua tan bien como lo hacía el ordenador. (Pasemos por alto el problema del tiempo de respuesta). Cada vez que Searle recibiera un texto escrito en chino, aplicaría las reglas y escribiría

los signos correspondientes a la respuesta que habría dado el ordenador.

Pero el caso de Searle es diferente al del ordenador: él sabe que no sabe chino y por lo tanto es consciente de que, al hacer el papel del ordenador, no se ha enterado de una palabra de la conversación que ha tenido con la mujer, aunque esa conversación haya sido coherente y capaz de engañarla, haciéndola pensar que estuvo dialogando con un ser humano que conoce la lengua china. La cuestión clave, por lo tanto, es la siguiente: ¿entiende el ordenador la conversación que ha mantenido con la mujer? Si no la entiende (igual que Searle no la entiende), ¿es el ordenador consciente de que no la entiende? Porque Searle sí lo es.

Luego no basta que un ordenador sea capaz de pasar la prueba de Turing para que podamos considerarlo tan inteligente como nosotros. Hacen falta dos cosas más: que el ordenador comprenda lo que escribe, y que sea consciente de la situación. Mientras eso no ocurra, no podremos hablar estrictamente de *inteligencia artificial*.

Searle propuso distinguir dos tipos de inteligencia artificial:

- Inteligencia artificial débil: la que podría alcanzar una máquina que pasara satisfactoriamente la prueba de Turing.
- Inteligencia artificial fuerte: la que tendría una máquina que tuviera una mente semejante a la humana, capaz de comprender y de saber si comprende o no comprende.

El problema es importante, porque introduce cuestiones para las que no tenemos respuesta, como si la consciencia puede programarse, la dualidad mente-cuerpo, cómo nos identificamos con los demás, o qué significa *comprender* un texto escrito o cualquier otra representación simbólica. Mientras todas estas cosas no se aclaren (si es que es posible aclararlas), es inútil pronosticar que la inteligencia artificial (la de verdad) está a la vuelta de la esquina, como hace Ray Kurzweil (Kurzweil 2014), quien tiene la esperanza de que dentro de pocos años será factible descargar la consciencia y la memoria de un ser humano en la memoria de un ordenador, lo que permitiría vencer a la muerte. A pesar de Kurzweil y en mi opinión, la probabilidad de que tal cosa ocurra, no ya durante los próximos 20 años, como él espera, sino alguna vez en el futuro, es prácticamente nula.

Referencias

Ackerman, E. 2014. "Can Winograd Schemas Replace Turing Test for Defining Human-Level AI?" Accessed August 11, 2014. http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/winograd-schemas-replace-turing-test-for-defining-humanlevel-artificial-intelligence/?utm_source=roboticsnews&utm_medium=email&utm_c.

Kurzweil, R. 2014. "Voz 'Ray Kurzweil'." Accessed August 11, 2014. http://es.wikipedia.org/wiki/Raymond_Kurzweil.

Santini, S. 2012. "El test de Turing y la inteligencia humana." 8 de noviembre de 2012. Accessed August 11, 2014. http://blogs.elpais.com/turing/2012/11/el-test-de-turing-y-la-inteligencia-humana.html.