Inteligencia Artificial y conciencia.

José Antonio Malpica Velasco.
Profesor Titular en el Departamento de Matemáticas de la UAH, en el área de conocimiento Ingeniería Cartográfica, Geodesia y Fotogrametría.

1 Introducción

La Inteligencia Artificial (IA) es la rama de las Ciencias de la Computación que estudia el software y hardware necesarios para simular el comportamiento y comprensión humanos. El objetivo último de la IA es simular la inteligencia humana en una máquina creando robots que sean conscientes y con sentimientos reales, similares a los humanos. Uno de los problemas más difíciles es la simulación de la conciencia, cualidad humana que hace que nos demos cuenta de nuestra propia existencia.

En la película de Kubrick del año 1968 "2001. Una Odisea en el Espacio" aparece un ordenador muy inteligente llamado HAL, viajando a bordo de la nave Discovery, con destino hacia una estación espacial; cuando HAL adivina las intenciones de la tripulación de desconectarlo, siendo consciente de lo que eso supone, surge dentro de él un impulso a seguir existiendo que le lleva a tomar la iniciativa e ir matando a los astronautas uno a uno. Esta película y otras por el estilo, junto a novelas de ciencia ficción, promocionaron durante años el interés por la IA. Con el avance del hardware y software en la década de los ochenta la investigación en IA pretendían simular capacidades humanas como la visión o el razonamiento. Los países desarrollados dedicaron gran cantidad de medios materiales y humanos a la IA; en aquellos momentos la euforia era generalizada y se pensaba que no se tardaría mucho en construir un ordenador como HAL; obviamente, no en su faceta psicópata de la ficción sino como consecuencia de aceptar el reto de construir un ser inteligente y consciente. En amplios círculos científicos se daba por hecho que todo fenómeno era posible expresarlo computacionalmente, corriente de investigación que se llamará más tarde IA fuerte; esta defendía que los ordenadores superarían a los humanos en todos los órdenes, y que la conciencia tenía carácter computacional. Es en este contexto cuando en 1989, el profesor de matemáticas de la Universidad de Oxford Roger Penrose [1], publica su libro "La nueva mente del emperador" que denuncia al estilo del famoso cuento del niño que anuncia que "el emperador está desnudo" toda esta contemporización alrededor de la IA fuerte. Ante el aluvión de críticas que recibió el libro, unos años más

tarde Penrose publicó un segundo libro [2] "Las sombras de la mente" para desarrollar en más detalle algunos de los puntos que levantaron más polémica. Es así como en la década de los noventa la IA dio marcha atrás, pues los científicos se dieron cuenta que los objetivos iniciales habían sido excesivamente ambiciosos, y los resultados prácticos no avalaban los teóricos, que en muchos casos se habían convertido en puras especulaciones. Se concentraron entonces en el desarrollo de hardware y software que abordasen tareas y problemas más modestos y puntuales. Hoy los ordenadores son muy eficaces resolviendo ciertos problemas como la toma de decisiones en asuntos comerciales, venciendo a los humanos en juegos como el del ajedrez, y en general ayudando al experto humano con sistemas expertos en una amplia variedad de disciplinas. Sin embargo, los científicos no han conseguido, después de unos treinta años de IA, simular en el ordenador comportamientos que resultan relativamente sencillos para los humanos, como por ejemplo las intuiciones. La simulación de la conciencia permanece hoy en día todavía muy en los principios, el año 2001 pasó y HAL permanecerá como ente de ficción durante muchos años todavía. La pregunta natural que surge ahora es: ¿será posible alguna vez crear un ordenador como HAL, que tenga conciencia? El debate en la actualidad se sitúa tanto en el plano filosófico como en el científico o tecnológico.

Repasemos las posturas de la IA ante la posibilidad de simular la conciencia por un ordenador, todavía hoy se encuentran defensores de cada una de ellas.

1 Fuerte IA: Como antes se ha dicho, los partidarios de esta postura piensan que toda actividad mental es de tipo computacional, incluidos los sentimientos y la conciencia, y por tanto se pueden obtener por simple computación.

2 Débil IA Cree que la conciencia es una característica propia del cerebro. Y mientras toda propiedad física se puede simular computacionalmente, no se puede llegar por este procedimiento al fenómeno de la conciencia en su sentido más genuino. Los que pertenecen a este grupo dicen que la simulación por ordenador de un huracán no es en sí mismo un huracán. O que la simulación de la digestión por el estómago no digiere nada. Se trata de un proceso no causal.

3 Nueva Física Esta postura defiende que es necesaria una nueva Física para explicar la mente humana y que quizás en el futuro se pueda simular, pero ciertamente no por métodos computacionales; para ello es necesario

que en el futuro se descubran nuevos métodos científicos que todavía se desconocen.

4 Mística Esta postura defiende que la conciencia no se puede explicar ni física, ni computacionalmente, ni por otro medio científico. Es algo totalmente fuera de la esfera científica, pertenece al mundo espiritual y no puede ser estudiada utilizando la razón científica, escapa al método de conocimiento racional heredado de la cultura griega.

Los que se sitúan en las dos primeras posturas están convencidos de que la conciencia es un proceso físico que emerge del cerebro, explicable con la ciencia actual, lo que sucede es que todavía no se ha llegado a desentrañar el misterio. Veremos en los apartados siguientes la diferencia entre las posturas de la IA fuerte y la IA débil. También veremos el ejemplo de la Habitación China de Searle que ofrece una analogía para mostrar cómo no es posible crear consciencia por simple simulación.

La postura tres es defendida especialmente por Penrose que declara que el cerebro posee propiedades que no son computacionales. Penrose se basa en el teorema de Gödel y la tesis de Church-Turing para afirmar que siempre habrá hechos verdaderos no alcanzables por un sistema computacional.

La postura cuatro puede parecer la más cercana a la postura teísta, por su consistencia con la mente universal en el budismo o por su coherencia con el dualismo (materia y espíritu) del cristianismo; sin embargo, los que piensen que solo la postura cuatro es compatible con la postura teísta se equivocan, cualquiera de las cuatro posturas sería compatible con la religión. La ciencia podrá o no responder a la pregunta de cómo funciona la mente, pero nunca responderá a la pregunta de por qué la mente existe o cuál es su fin último. Son ámbitos distintos de trabajo aunque puedan existir algunas interrelaciones.

2 Computabilidad e Indecidibilidad

En el año 1900 el prestigioso matemático alemán David Hilbert propuso una serie de problemas, cuya resolución se consideraba clave para el avance de las matemáticas y la ciencia en general, entre ellos estaba el conocido con el Entscheidungsproblem, que se pregunta si existe un procedimiento

mecánico que resuelva todos los problemas matemáticos pertenecientes a un tipo dado bien definido. Este problema fue el que motivó a Turing a proponer la famosa máquina que lleva su nombre.

La máquina de Turing

Alan Turing fue un matemático británico especializado en criptografía que jugó un papel importante dentro del grupo que descifró los códigos secretos alemanes durante la Segunda Guerra Mundial. Turing concibe su máquina como una descripción idealizada de la actividad que llevaban a cabo los contables realizando tediosos y rutinarios cálculos con lápiz y papel ayudados a veces de tablas trigonométricas o de logaritmos. Hasta el advenimiento de los ordenadores este cálculo numérico contable empleaba a miles de personas en bancos, administración pública y en muchas otras empresas, era una actividad muy común en aquellos días; el mismo Turing lo indica en una de sus publicaciones [3] "si a una persona se le facilita papel, lápiz y goma de borrar y se le somete a una disciplina estricta, dicha persona se convierte en lo que se denomina una máquina universal". Wittgenstein fue más radical, si cabe, a la hora de referirse a las máquinas de Turing: "Estas máquinas son humanos que calculan."

Una máquina de Turing es un modelo formal de ordenador, un concepto teórico de computación que formaliza el concepto de algoritmo. Más adelante se dará una definición más formal de algoritmo, de momento puede servir, la siguiente definición: un algoritmo es todo procedimiento que realiza una tarea determinada en un número finito de pasos. Una máquina de Turing consiste en una cinta de longitud infinita, dividida en celdillas, donde cada una contiene una letra de un alfabeto o está en blanco. El alfabeto consiste en un conjunto finito de símbolos, incluido el espacio en blanco. La máquina tiene un cabezal para leer los símbolos en las celdillas y escribir nuevos símbolos que sustituyan a los existentes cuando corresponda; cuando el cabezal ha terminado de leer o escribir en una celdilla se mueve a la derecha o a la izquierda de esa celdilla, una sola posición cada vez, salvo que se llegue al estado de parada. La máquina tiene un número finito de estados, y se dice que siempre se encuentra en cierto estado. Un programa consiste en una lista de instrucciones, cada instrucción le indicará a la máquina una computación a realizar; es decir, una serie de acciones, dados un estado actual y un símbolo determinado bajo el cabezal, la instrucción le indicará qué símbolo debe ser escrito en la cinta, y si el cabezal debe moverse un paso a la izquierda o uno a la derecha. Una instrucción viene definida por una quíntupla como la siguiente:

(estado inicial, valor inicial, nuevo estado, nuevo valor, movimiento)

La cinta se utiliza para almacenar datos, pero también se puede utilizar para guardar una serie de instrucciones (pequeños programas o subprogramas). En este último caso se dice que la máquina de Turing emula a otra, la que se encuentra en la cinta, este tipo de máquina se conoce con el nombre de máquina universal. Una observación crucial es que se demuestra que solo hay un número contable de máquinas de Turing. Entendiendo por conjunto contable aquel que es finito o se puede poner en correspondencia uno a uno con el conjunto de los números naturales.

En Malpica [4] se pude ver un ejemplo de máquina de Turing que decide si la cadena de entrada que se le facilita es un palíndromo o no. Más ejemplos de máquinas se pueden encontrar en la siguiente página web:
http://wap03.informatik.fh-wiesbaden.de/weber1/turing/tm.html

Tesis Church-Turing

El nombre algoritmo, y los adjetivos computable, mecánico y recursivo se utilizan todos para denotar el carácter propio de las operaciones que puede realizar una máquina de Turing. Definamos lo que es un procedimiento mecánico M:

1.- M se expresa mediante un número finito de instrucciones, donde cada instrucción se construye a partir de un número finito de símbolos.
2.- M producirá, si se ejecuta sin error, el resultado deseado en un número finito de pasos.
3.- Un ser humano puede ejecutar M (en la práctica o en teoría) sin necesidad de utilizar ninguna máquina o artilugio, sólo provisto de lápiz y papel.
4.- El humano que ejecutase el procedimiento M simplemente tiene que seguir las instrucciones y los pasos que definen a M, no necesita de intuiciones o comprensión de lo que se está haciendo.

Para cualquier procedimiento mecánico se puede encontrar una máquina de Turing que lo represente y ejecute. En este sentido nos preguntamos si el concepto de máquina de Turing incorpora todo procedimiento M mecánico. Casi paralelamente al trabajo de Turing, sólo un poco antes Alonzo Church descubrió el cálculo landa para abordar el mismo problema propuesto por Hilbert en 1900. Un poco más tarde estos mismos autores demostrarán que

el cálculo landa y la máquina de Turing son equivalentes. Esto vino a conocerse como la tesis de Church-Turing y viene a decir, que la máquina de Turing (y el cálculo landa) definen lo que se entiende por algoritmo o proceso mecánico. Esta tesis no era tan evidente hace 70 años, como lo puede ser ahora gracias a la capacidad computacional de los ordenadores de hoy día.

Turing demostró que hay cierta clase de problemas que no tienen solución algorítmica, entre ellos el más famoso es "El problema de la parada". Turing muestra que para cierta clase de problemas la máquina que lleva su nombre no para; es decir, no decide sobre los problemas, es lo que llamamos la cuestión de la indecidibilidad. Una demostración simplificada y divulgativa del problema de la parada puede verse en [4]. Con la indecidibililidad el problema propuesto por Hilbert queda resuelto con una respuesta negativa: No puede haber un algoritmo general que resuelva todo problema matemático. Hay que tener cuidado de no extender las conclusiones más allá de lo que permite el mismo problema resuelto. Los algoritmos no dicen, ni pueden decir por sí mismos, nada sobre las verdades de las sentencias o proposiciones matemáticas. Es indiscutible que la validez de un algoritmo debe establecerse por medios externos al propio algoritmo. Menos aún se pueden sacar conclusiones que impliquen a sistemas físicos como el cerebro. De momento solo se puede decir que existen problemas para los que no existe una máquina de Turing mientras esos mismos problemas sí se intuyen por la mente humana. O de otra forma hay procesos que puede realizar la mente humana que no son algorítmicos en el sentido de existencia de una máquina de Turing.

3 Demostrabilidad e Incompletitud

Teorema de Gödel

La demostración del teorema de Gödel en su forma original es muy complicada y detallista, como no podía ser menos tratando con sistemas formales; sin embargo, si nos permitimos ciertas libertades en lo referente al rigor lógico es posible expresar de una manera más fácil la estrategia e ideas en que se fundamenta, esto es lo que intentaremos a continuación:

Como en cualquier sistema formal consideramos un alfabeto de símbolos, reglas que combinen los símbolos y formen las proposiciones, un conjunto de

axiomas y un aparato lógico de reglas que permita obtener nuevas proposiciones a partir de otras ya obtenidas de los axiomas.

Gödel asigna un número a cada símbolo primitivo del alfabeto. Una vez que a cada símbolo le corresponde un número se continúa estableciendo una regla que asigne un número a cada proposición. Finalmente se define una regla para asignar un número a cada cadena de proposiciones en sucesión lógica que produce cada nueva proposición, i.e. se asigna un número a cada demostración. En lo que se conoce como la numeración de Gödel. Una vez que cada proposición del sistema formal tiene asignado un número estamos en disposición de analizar la estructura relacional entre dos proposiciones cualesquiera analizando la relación aritmética entre sus números correspondientes. El sistema de codificación de Gödel utiliza el producto exponencial de números primos, y se basa en el teorema de factorización de números primos que dice que todo número se puede expresar de forma única como producto de números primos.

Una proposición será consecuencia de otra si sus correspondientes números están relacionados de una manera determinada; así, Gödel utiliza la aritmética de los números enteros para establecer un metalenguaje del sistema formal. Encontrará así una proposición que es verdadera pero que no es demostrable; véase Goldstein [5] para más detalles, allí se ofrece una demostración del teorema Gödel en forma divulgativa. Este teorema es conocido como teorema de incompletitud de Gödel; demuestra Gödel que hay al menos una verdad que no es demostrables dentro del sistema. Se podría pensar entonces en extender el conjunto de axiomas con esta verdad y resolver el problema; sin embargo, el problema permanecería pues siguiendo el mismo razonamiento existiría una nueva verdad que no sería demostrable en el sistema, y así hasta el infinito. En esto consiste la incompletitud de un sistema formal. A partir de este teorema es posible deducir otro que se conoce como el teorema de inconsistencia y que dice que en todo sistema formal del que forme parte la aritmética no se puede demostrar su propia consistencia, i.e. no se puede estar seguro de encontrar alguna vez una proposición que sea verdadera y falsa a la vez.

La relación entre sistemas formales y computabilidad es estrecha, los cálculos que se realizan desde los axiomas hasta alcanzar la verdad o falsedad de la fórmula inicialmente propuesta es un proceso computacional, en el sentido indicado de algoritmo de la sección anterior. Así de alguna manera se puede decir que demostrabilidad es equivalente a computabilidad y que completitud es equivalente a decidibilidad.

Las implicaciones de la incompletitud en la propuesta de Gödel desmoronan el paraíso de los mecanicistas de encontrar un sistema formal en el que se pudiera demostrar toda verdad matemática, es decir reducir la matemática a un sistema lógico formal universal. Parece lógico extender el teorema de Gödel al ámbito de la física, esa física cuidadosamente axiomatizada en el siglo XIX, especialmente la mecánica, que se apoya en la matemática. Pero pensando con más cuidado pues no es tan inmediato como parece, el teorema de Gödel debe entenderse en su justa medida, en cuanto que es un teorema lógico matemático dentro del mundo de la ideas abstractas, extender su conclusiones al mundo de las realidades físicas supone un salto cualitativo. Penrose consciente de este hecho utiliza la mecánica cuántica para mostrar procesos físicos que se dan en el cerebro que no pueden conducir a un mecanicismo. Sobre las implicaciones en el mundo real del teorema de Gödel se ha escrito mucho desde distintas perspectivas y es todavía hoy un tema de discusión. De todas formas, desde un sentido puramente práctico, en física se construye un sistema y cuando se encuentra mediante experimentos que algún fenómeno que no encaja dentro del sistema se reconstruye la teoría [6], proponiendo un nuevo sistema.

4 ¿Qué tipo de inteligencia y conciencia posee una máquina?

Al tratar de contestar esta pregunta se da otro salto cualitativo, del ámbito lógico matemático al ámbito científico filosófico. El primero es mas restringido y por tanto las reglas de "juego" están bien definidas, como contrapartida tampoco se puede decir mucho sobre problemas de alcance más amplio que resultan más interesantes. Turing también dio ese salto y propuso una prueba para determinar cuando se puede considerar inteligente una máquina. Así establece que una máquina se puede considerar inteligente si al comunicarse con un humano (que no ve físicamente a la máquina) este no sabría decir si está hablando con una máquina o con otro humano como él. Este test aunque no lo ha pasado ninguna máquina todavía, sí que parece alcanzable en un futuro no lejano. Ello animó a los seguidores de la IA fuerte a aventurar el advenimiento de máquinas inteligentes que alcanzarían las mismas características de la inteligencia humana; sin embargo, Searle presentó (Searle, 1980) el siguiente argumento en contra: Supongamos una persona que sepa solo inglés y se le encierra solo en una habitación, la puerta dispone de una pequeña rendija que le sirve de comunicación con el exterior. La comunicación se establece exclusivamente mediante hojas de

papel escritas en chino que se pasan por la rendija. La tarea de esta persona consiste en seguir unas instrucciones escritas en inglés, que se encuentran en la habitación, de cómo manipular caracteres chinos. Cuando se le pasan desde el exterior preguntas en chino, esta persona utilizando las instrucciones en inglés contestaría también en chino, sin saber el significado real de lo que le preguntan y de lo que él contesta; sin embargo para observadores que se encuentran en el exterior daría la sensación de que sabe chino.

Con esto Searle muestra que aunque veamos hablar a un ordenador y contestar a nuestras preguntas, en realidad no son capaces de entender la lengua. Los ordenadores utilizan los elementos sintácticos de la lengua pero no los semánticos.

El argumento de la Habitación China no va en contra de la IA débil, no pretende demostrar que los ordenadores no puedan pensar. De hecho, Searle piensa que el cerebro es una máquina que piensa. En realidad, va dirigido contra la IA fuerte que mantiene que la computación formal con símbolos puede producir pensamiento.

Con todo es conveniente distinguir entre epistemología y ontología, es decir, entre "¿Cómo conozco?" y "Qué es lo que conozco cuando conozco? El test de Turing cae en este error de mezclar epistemología y ontología y de ahí viene la confusión que se ha dado incluso en las discusiones entre expertos. El ámbito epistemológico se relaciona mejor con la ciencia que el ontológico. El ontológico sobrepasa al científico, porque la ciencia carece de las herramientas para contestar a muchos de los problemas que plantea la filosofía, que sin embargo resultan más radicales para el ser humano.

Es importante separar y definir específicamente qué se entiende por conciencia. Si no, se podría argumentar que un simple termostato es consciente de que hace frío en la habitación en que está instalado, porque se dispara y pone en funcionamiento la calefacción cuando en la habitación se alcanzada una cierta temperatura. Sólo nos vamos a referir a niveles de consciencia superior, cualidad propia de los seres humanos; es decir, definimos conciencia como la capacidad para debatir, reflexionar y darse cuenta de la propia existencia. Penrose considera que existen intuiciones de la mente que no se pueden expresar algorítmicamente, y apoyado en el teorema de Gödel dice que la mente no se puede simular con la ciencia que ahora conocemos, indica que tiene que surgir un nuevo paradigma científico para que se pueda abordar este problema.

Todo esto entronca con el ámbito filosófico del mundo de las ideas de Platón. Nos movemos así entre las matemáticas y la filosofía, lo que Einstein consideraba los problemas genuinamente interesantes. Gödel era un platonista entre positivistas en el círculo de Viena, él estaba callado mientras los otros hablaban de la obra de Witteinstein. Gödel en lugar de hablar ofrece una demostración que lleva directamente al mundo de las ideas de Platón, y que también dice algo en contra de "el hombre es la medida de todas las cosas" de los positivistas del círculo de Viena [5].

Se podrán hacer muchas cosas en inteligencia artificial. Pero cuando en los siglos XVII y XVIII se intentaba construir una máquina que volara se miraba a las aves y como simular su capacidad de volar, se intentaba ponerles alas y una cola que se movieran de forma similar a las de las aves. A principios del siglo XX surge el avión que volaba pero que se parecía a las aves solo en las alas y la cola. Aunque los aviones no son como la aves sí son mas útiles a los humanos que las aves en el sentido puro de utilidad, pues los aviones nos permiten viajar rápidamente de un punto a otro, transportando toneladas de material, cosa que no se lograría ni con las aves mitológicas; sin embargo las aves aventajan a los aviones en que son capaces de reproducirse por si mismas; es obvio que el sistema de reproducción de las aves aventaja al de los aviones, al menos en economía, los aviones son incapaces de poner huevos por ahora. Sirva esta analogía para indicar lo que cabe esperar de sistemas desarrollados intentando simular la inteligencia y la conciencia; probablemente se consigan en un futuro androides, seres artificiales, muy inteligentes, incluso más que los humanos en la realización de ciertas tareas, pero obviamente siempre serán distintos de los humanos. Por el simple hecho de haber sido creados por el hombre, a los androides no se les podrán atribuir cualidades humanas.

En general, los seres vivos no son como las máquinas hechas por los humanos, como les gusta decir a algunos mecanicistas. Un androide o un robot es como un coche, se fabrican las piezas independientemente: las ruedas, el volante, los cilindros del motor, etc. que luego se ensamblan entorno a una unidad central que puede ser el ordenador de abordo. El ensamblaje se realiza de fuera a dentro. En los seres vivos el proceso se da al revés de unas pocas células se desarrolla todo el ser vivo. El ensamblaje se realiza de dentro a fuera. Así Borden [8] argumenta que el metabolismo bioquímico es necesario para la vida. Según esta autora puede ser posible producir vida artificial pero probablemente tendrá que ser a través de procesos bioquímicos y no metálicos electrónicos. Un programa informático,

esté o no insertado en un robot, no tiene intencionalidad, no tiene conciencia y, por tanto, no puede tener una semántica intrínseca.

Bibliografía

[1] Penrose R. "The Emperor's New Mind. Concerning Computers, Minds, and the Laws of Physics". Vintage 1989. (existe traducción al castellano de esta obra en la editorial Mondadori de 1991).

[2] Penrose R. "Shadows of the Mind: A Search for the Missing Science of Consciousness" Oxford University Press 1994. (traducción al castellano en la editorial Crítica de 1996).

[3] Turing A., (1948) Intelligent Machinery. National Physical Laboratory Report. En Meltzer, B., Michie, D. (eds) 1969. Machine Intelligence 5. Edibrgh University Press.

[4] Malpica J.A, "Introducción a la Teoría de Autómatas" UAH 1998.

[5] Goldstein R. Incompleteness. The Prof. and paradox of Kurt Gödel. (2005) Atlas Books. New York.

[6] Barrow J. D. Impossibility. The Limits of Science and the Science of Limits" Oxford University Press 1998. (traducción al castellano en Gedisa 1999)

[7] Searle J. (1980). Brains and Programs, Behavioral and Brain Sciences, 3:417-57 (disponible en internet)
(disponible en internet: www.AlanTuring.net/intelligent_machinery).

[8] Boden, M. A. (1999). Is metabolism necessary? Brittish Journal of the Philosophy of Science, 50(2), 231–248.