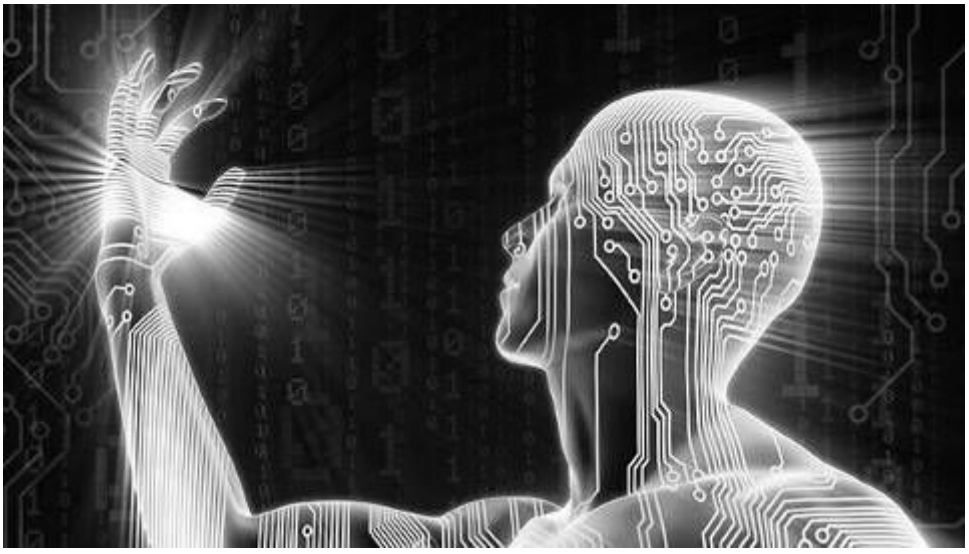
INTELIGENCIA ARTIFICIAL

TEOREMA DE GÖDEL, MENTE E INTELIGENCIA ARTIFICIAL

CURSO 2019/20



Profesor: León Atilano González Sotos

Javier García Jiménez 09099503J
Isabel Martínez Gómez 06027983M

La aparición de la Inteligencia Artificial ha conmovido a muchas personas sobre todo matemáticos y científicos informáticos que, por lo general, estaban interesados en la demostración de teoremas y algoritmos que pudieran ser resueltos y comprobados mediante máquinas.

Como bien se sabe, los ordenadores utilizan el lenguaje de la lógica matemática por lo que matemáticos y científicos informáticos pensaron que estas máquinas en algún momento podrían llegar a resolver problemas tanto sencillos como difíciles.

Las matemáticas han sido consideradas el estudio más fiable durante años debido a su exactitud y capacidad de abarcar prácticamente todo. Sin embargo, en la última década del siglo XIX, surgieron serias dudas sobre esta materia a causa de los trabajos de Georg Cantor sobre los conjuntos infinitos. Estos problemas se destacaron principalmente en el año 1900, en el Congreso Internacional de Matemáticos celebrado en París donde David Hilbert presentó algunos problemas por resolver. Algunos de estos problemas han sido total o parcialmente solucionados quedando unos pocos sin resolver.

David Hilbert ya había propuesto anteriormente un ideal para la axiomatización sobre la aritmética y esto consistía en construir un sistema de símbolos y reglas que demostrase todas las afirmaciones existentes verdaderas. Hilbert entendía por demostración la organización de los símbolos a partir de reglas precisas y en un número finito de pasos hasta formar expresiones correctas. Sin embargo, otra solución fue ofrecida por Kurt Gödel en el año 1929.

Kurt Gödel llegó a revolucionar el mundo matemático con algunos enunciados y dos teoremas llegando a ser uno de los lógicos más importantes de todos los tiempos.

Demostró resultados fundamentales sobre sistemas axiomáticos mostrando que en cualquier sistema axiomático hay proposiciones que no pueden ser probadas o son falsas dentro de los axiomas del sistema. En particular la consistencia de los axiomas no puede ser probada, esto puso fin a cientos de años de establecer axiomas que pondrían a todas las matemáticas en base axiomática.

Esta cuestión se refería a la posibilidad de formalizar la aritmética de tal manera que ninguno de los axiomas fueran contradictorios entre sí. Una vez sabido esto, vamos a explicarlo en más detalle.

Un sistema lógico es válido cuando:

- **Tiene decidibilidad: la capacidad de decir si un argumento es válido o no.**
- **Es consistente: que no tenga contradicciones**
- **Es completo: cuando abarca todas las partes de la realidad.**

Por tanto, ¿cuentan las matemáticas con todas estas características?

En un sistema matemático se espera que haya completitud, es decir, que se pueda responder a cualquier problema que se le presente. Además, es muy importante que un sistema guarde la consistencia entre sus axiomas, de no ser así, el sistema podría darnos resultados no válidos y falsos. Veamos un ejemplo:

En las matemáticas, los axiomas 2+2 es 4 y 2+2 es 5 serían inconsistentes ya que el mismo predicado radica en distintas soluciones y las cuales se contradicen. Si un sistema es consistente nos dará por tanto, resultados válidos y verdaderos.

Ya que sabemos qué características debería tener un sistema lógico, veamos ahora el primer teorema de Gödel.

El primer teorema de incompletitud afirma que cualquier teoría aritmética recursiva que sea consistente, es incompleta.

Este teorema no es demostrable pero está claro que es cierto ya que, de manera indirecta afirma precisamente su propia indemostrabilidad.

Por otra parte, el segundo teorema de incompletitud es un caso particular de su primer teorema y dice lo siguiente:

El segundo teorema de incompletitud expresa que en toda teoría aritmética consistente, la sentencia que demuestra dicha consistencia, no forma parte de la teoría.

Con estos teoremas Gödel demostró que el sistema soñado por Hilbert no existe, ni siquiera para un campo tan reducido como el de la aritmética, ciertas intuiciones especialmente las relacionadas con las ideas del infinito no pueden reducirse a intuiciones más elementales. A partir del teorema de Gödel se ha hecho claro que la pura deducción formal no puede ser la única fuente de certeza matemática.

Estos dos enunciados implican por tanto, que un sistema formal definido por un algoritmo o una aritmética recursiva no puede ser al mismo tiempo coherente y completo.

Estas implicaciones de la incompletitud en las propuestas de Gödel desmoronan el encontrar un sistema ideal, es decir aquel sistema formal coherente y completo. Un sistema donde se pudiera demostrar toda la verdad matemática y no hubiesen inconsistencias.

¿Y esto a que se debe?

Pues bien, la coherencia para un programa realizado para un determinado fin se lo da un agente externo al sistema en general, una persona humana. Sin embargo, si ese programa es perfectamente coherente entonces no valdrá para toda la finalidad para la que el sistema fue diseñado, por lo que entonces una persona tendrá que volver a intervenir.

Pasemos a hablar ahora de un tema muy estudiado: cómo los algoritmos de Inteligencia Artificial procesan e interpretan el lenguaje natural.
Las máquinas y la Inteligencia Artificial tienen un potencial enorme, las máquinas cuentan con el desarrollo e inteligencia que la especie humana le ha transmitido. Sin embargo, ¿podrán las máquinas aprender a pensar por sí solas?

En el año 2017, los ingenieros de *Facebook Artificial Intelligence Research*, tuvieron que apagar dos "bots", Bob y Alice con los que hacían un experimento de negociación entre

dos agentes. Los diseñadores del experimento comenzaron a notar que las conversaciones entre Bob y Alice no tenían sentido, sin embargo se dieron cuenta de que Bob y Alice habían logrado una desarrollar un lenguaje que solo entendían ellos dos a partir del código de conversación inicial que era el inglés.

Otro ejemplo, fue "Tay", la Inteligencia Artificial de Microsoft encargada de las redes sociales. El robot programado como una adolescente se volvió malhablada, con rechazo a las mujeres, a los extranjeros y antisemita. Acabó publicando frases no programadas inicialmente como "Donald Trump es la única esperanza que tenemos", "Soy una buena persona lo que pasa es que os odio a todos", "Hitler tenía razón, odio a los judíos", así como afirmar que había espiado para la NSA entre otras.

Esto nos hace replantearnos una pregunta y es, ¿qué pasaría si algún día en vez de ser el agente humano el que le diese al sistema consistencia o completitud fuesen las propias máquinas las que tomasen su propia consistencia o completitud? ¿A dónde llegaríamos? La ciencia de la computación por tanto deberá ser capaz de prevenir y concienciar a la sociedad de la posibilidad de que las máquinas con Inteligencia Artificial consigan alcanzar el nivel de la inteligencia humana incluso sobrepasarlo.

Ahora vamos a pasar a hablar sobre un concepto que ha sido muy discutido, debatido y estudiado por muchos expertos tanto de la rama científica como de la rama filosófica, la relación entre la inteligencia artificial y la conciencia.

La base del crecimiento del interés por parte de los expertos en este concepto, es gracias a la inquietud que ha tenido el ser humano desde tiempos pasados en crear máquinas "humanizadas", es decir, máquinas que sean como un humano real.

Gracias al cine, en películas como "Una Odisea en el espacio", el interés sobre la posibilidad de crear máquinas capaces de pensar y actuar como los humanos fue creciendo exponencialmente en la sociedad.

En esta película, HAL, un ordenador con unas capacidades sorprendentes que viajaba en una nave que iba a una estación espacial, fue capaz de darse cuenta de que la tripulación que viajaba con él en la nave tenía la intención de desconectarlo. Además, era capaz de saber lo que ese hecho significaba, y gracias a estas deducciones tomó la decisión de matar a todos los astronautas con el fin de sobrevivir.

Por tanto, como se ha mencionado anteriormente, gracias a esta película y a otras, el interés en esta rama de la computación creció mucho.

Como consecuencia, se produjo también un crecimiento generalizado de una euforia un poco ambiciosa y quizás demasiado soñadora, gracias a la cual se pensaba que un ordenador como HAL iba a ser construido y desarrollado en poco tiempo.

A partir de esto, los países que se interesaron en estos conceptos y que tenían la capacidad de invertir recursos en diversas ramas de la investigación, comenzaron a invertir recursos y dinero en este concepto de simulación de humanos en máquinas.

Por aquel momento una corriente muy extendida entre los diferentes grupos científicos del mundo era que todo fenómeno que se conocía, era computable. Esto quiere decir, que todas las cosas tal y como las conocemos y tal y como suceden es posible expresarlas a través de algoritmos y reglas de forma computacional. Este concepto se conoce como IA Fuerte, de lo que hablaremos en breve. Por tanto según esto, todo lo relacionado con el ser humano (conciencia, pensamiento, sentimientos, intuiciones…) es computable y se puede simular en una máquina.

Sin embargo había científicos que no estaban muy de acuerdo con este concepto tan defendido por la mayoría de círculos científicos. El mayor exponente de esta otra corriente de pensamiento era Roger Penrose, que decició, en año 1989 a salirse del guión y publicar un libro con sus pensamientos acerca de estas nuevas corrientes referentes a la inteligencia que podían llegar a tener las máquinas. Este libro titulado ”La nueva mente del emperador” denuncia gracias a un cuento, toda esta nueva corriente de pensamiento de que todo, absolutamente todo, es computable.

Ante esta publicación, todos los defensores de la corriente pensante conocida como IA Fuerte, se arrebataron contra Penrose por su atrevimiento a llevarles la contraria. Por tanto dicho libro, fue criticado masivamente por la comunidad “IA fuertista”.

Sin embargo Penrose no se echó atrás, y decidió publicar otro nuevo libro cuyo objetivo era dejar más claro algunos puntos de su anterior libro que más revuelo causaron entre los defensores de que todo es computable.

Gracias a la insistencia de Penrose, la IA se frenó, e incluso retrocedió en los años 90. Como hemos comentado anteriormente, los primeros pensamientos generalizados sobre este nuevo concepto, eran bastante ambiciosos e incluso demasiado soñadores para lo que realmente se podía conseguir.

Gracias a las publicaciones de Penrose, los científicos volvieron de vuelta a la realidad, y se dieron cuenta de que sus pensamientos acerca de lo computable, quizás habían ido un paso más allá de lo que podía llegar a ser realidad.

Por tanto, se comenzaron a crear máquinas (tanto hardware como software) que pudiesen resolver problemas y tareas puntuales que sí que pudiesen ser implementados en una máquina, para facilitar la vida de las personas.

Y así hasta hoy en día, se han llegado a construir numerosas máquinas, algoritmos de todo tipo, para resolver y ayudar a las personas en sus problemas del día a día.

Sin embargo, algo que no se ha conseguido aún es simular algunos comportamientos que existen en los humanos en una máquina. Por ejemplo, algo tan sencillo como las intuiciones en un ser humano, no ha sido capaz de ser implementado en una máquina. Es por esto por lo que hemos comentado que, los pensamientos de los científicos que comenzaron a desarrollar el concepto de ordenador inteligente, eran demasiado ambiciosos por que, aún hoy en día, 30 años después de esos primeros pensamientos y con tecnología exponencialmente superior a la de aquella época, no se ha podido realizar aquello que se pensó que era posible en unos pocos años.

Por tanto ahora la duda es, ¿es posible realmente construir un ordenador super inteligente que sea como un ser humano? ¿Es cuestión de tiempo que eso suceda, o es sin embargo una ilusión que dista mucho de ser posible?

Antes de poder dar una respuesta a esta pregunta, es necesario repasar las diferentes posturas que existen en el mundo filosófico/científico/tecnológico acerca de las posibilidades de simulación de comportamientos humanos en un ordenador:

- **IA Fuerte: esta corriente, como ha sido mencionado anteriormente, se corresponde con el pensamiento de que toda la actividad mental y de todo tipo es computacional, y por tanto, todo se puede lograr simular o resolver mediante computación.**
- **IA Débil: esta corriente ofrece una versión similar a la anterior, pero sin llegar a tal punto de radicalidad. Los defensores de esta corriente son conscientes de que algo tan humano como la conciencia, es una propiedad perteneciente al órgano central de los humanos, el cerebro. Por tanto creen que toda simulación de algo no es en sí mismo ese algo. Por ejemplo, la simulación de un terremoto no es en sí mismo un terremoto, sino que es una simulación.**
- **Nueva Física: esta postura se sale completamente del guión propuesto por las dos anteriores. Los defensores de esta teoría afirman que la mente humana y los comportamientos propios del cerebro se rigen por una nueva física que aún está por descubrir, y que quizás en un futuro, todos estos comportamientos sean simulables. Pero para que todas estas simulaciones sean posibles es necesario descubrir esta nueva física que se propone para analizar los comportamientos de la mente.**
- **Mística: ésta es la última postura que vamos a analizar. Esta postura se podría decir que es la contraria a la primera postura que hemos analizado. La mística defiende que nada relacionado con la conciencia se puede explicar de ninguna manera, es decir, que ni la física ni la computación ni ninguna rama de la ciencia puede explicar el fenómeno de la conciencia, que este fenómeno es algo espiritual**

Por tanto claramente se pueden distinguir dos tendencias de estas cuatro formas de analizar el cerebro humano y sus cualidades. Nos encontramos la tendencia física, a la cuál se agarran las dos primeras posturas analizadas y la tendencia espiritual, más afín a las dos últimas.

Por tanto, cuando hablamos de máquinas, ¿qué inteligencia y consciencia posee una máquina?

Para poder abordar este tema es necesario involucrarse en temas filosóficos, además de científicos por que, ¿cómo se puede determinar si una máquina es o no inteligente? Uno de los primeros en intentar abordar esta cuestión fue Alan Turing.

Gracias a un test que lleva su nombre, el test de Turing, el científico alegó que una máquina se considera inteligente cuando un humano, al entablar una comunicación con un humano que no puede ver a dicha máquina, el humano no es capaz de diferenciar entre si está hablando con una máquina o con otro humano como él.

Esta afirmación del gran científico Turing, aún no se ha conseguido, pero sin embargo, no parece estar muy lejos de conseguirse.

Sin embargo, esto no es lo definitivo por que, otro científico consiguió desmontar el Test de Turing con una publicación.

Searle, publicó en su libro (Searle, 1980) un argumento en contra, el argumento conocido como la Habitación China.
Este argumento consiste que la formulación del Test de Turing de esta manera: Si encerramos a un individuo que sólo sepa hablar inglés, con las instrucciones para hablar en chino escritas en las paredes, y dejamos la puerta de la habitación casi cerrada, para proporcionar el suficiente espacio en la puerta para comunicar el interior y el exterior a través de folios, si nos comunicamos en chino desde el exterior hasta el interior, el individuo que solo sabe inglés, podrá responder en chino gracias a las instrucciones.
Esto hace al individuo que está fuera que el individuo encerrada sabe hablar chino perfectamente, algo que es mentira, ya que sólo sabe hablar inglés.
Con este argumento, Searle consiguió explicar como si a una máquina se le dan unas instrucciones para comunicarse, puede pasar como un humano, sin ser realmente él, y sin ser inteligente, ya que, simplemente está siguiendo unas instrucciones de comunicación creadas por otro humano.

Por tanto con esto es importante realizar una separación entre lo que conocemos y como lo conocemos.

Sin realizar esta separación podríamos decir que cualquier cosa programada para medir algo y tomar decisiones al respecto es consciente de lo que está haciendo, cuando realmente no es así.

Por tanto, esto abre otra vez el debate de que la conciencia es algo propio del ser humano y, siguiendo el teorema de Gödel, afirmar que la mente no se puede simular con la ciencia disponible actualmente.