ChatGPT y el fin de un mundo

EPISODIO 2

¿QUÉ LÓGICA HAY TRAS UN LLM?

Tommaso Flaminio, Lluís Godó
IIIA – Instituto de Investigación en Inteligencia Artificial
CSIC – Consejo Superior de Investigaciones Científicas

*«Es importante tener presente que los modelos de lenguaje no incorporan ningún mecanismo que asegure la veracidad del contenido que producen:
las frases que generan pueden sonar totalmente plausibles y razonables, pero no hay garantía de que sean ciertas (a veces lo serán y otras veces no).»*

Raquel Fernández, episodio 1

¿Qué es la *lógica*?

La lógica es la disciplina que estudia los pasos necesarios para generar un razonamiento válido.

Para ello, proporciona principios y reglas, llamadas *reglas de inferencia*, para analizar la estructura de un argumento y asegurar conclusiones válidas a partir de un punto de partida inicial.

ESTAS SON ALGUNAS DE LAS REGLAS DE INFERENCIA DE LA LÓGICA CLÁSICA

Si es verdad que llueve, entonces no es verdad que no llueve.	$A \rightarrow \text{no}(\text{no } A)$
En cualquier momento, o llueve o no llueve.	$A \text{ o no } A$
Si llueve y, cuando llueve, se moja el suelo, entonces el suelo está mojado.	$(A \,\&\, A \rightarrow B) \rightarrow B$
Si el suelo no está mojado y, cuando llueve, se moja el suelo, entonces no llueve.	$(\text{no } B \,\&\, A \rightarrow B) \rightarrow \text{no } A$
Si, cuando llueve, se moja el suelo, y cuando se moja el suelo, sale Kim a jugar con los caracoles, entonces, cuando llueve, sale Kim a jugar con los caracoles.	$(A \rightarrow B \,\&\, B \rightarrow C) \rightarrow A \rightarrow C$
Si de lo que dice Kim se deduce que llueve y, también, que no llueve, entonces Kim miente.	$(K \rightarrow A \,\&\, K \rightarrow \text{no } A) \rightarrow \text{no } K$

¿Para qué sirven *las reglas de la lógica?*

La lógica no trata de distinguir lo verdadero de lo falso. Propone una serie de reglas para identificar cuándo un razonamiento está bien construido; y nos garantiza que, si damos por válidas las premisas del razonamiento, entonces la conclusión también será válida.

Podemos imaginar las reglas de la lógica como un mapa que nos ayuda a representar el mundo de forma estructurada y consistente.

¿Cómo establecemos *estas reglas?*

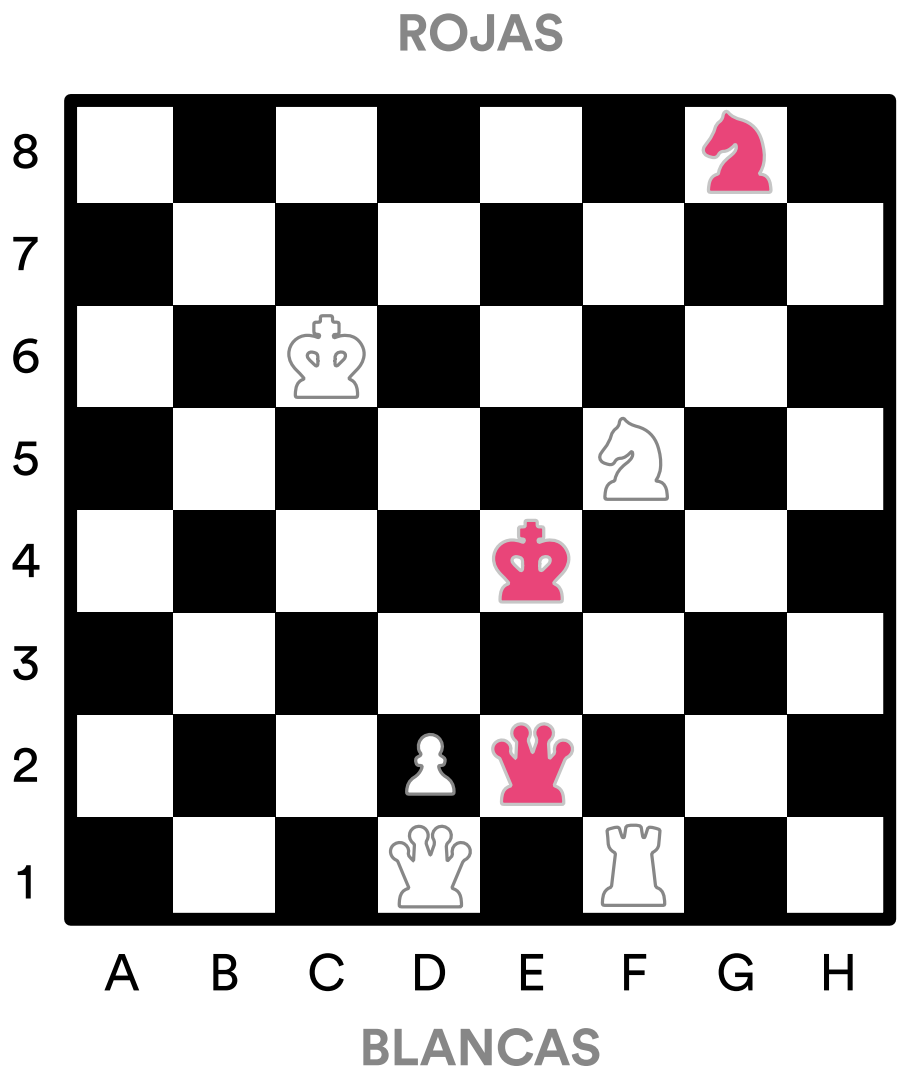
Primero, debemos elegir el ámbito sobre el que queremos pensar, el mundo en el que desarrollar nuestros razonamientos.

Por ejemplo, **el mundo de los números, el de las leyes, el de los juegos...**

Luego, construimos un lenguaje: un conjunto de palabras y símbolos que nos permiten representar y describir los objetos de este ámbito, así como sus propiedades fundamentales.

- Color blanco o rojo.
- Peón, reina, rey, torre, alfil…
- Enroque, jaque, horquilla…
- 1, 2, 3, 4, 5, 6, 7, 8.
- A, B, C, D, E, F, G, H.

Por último, definimos un conjunto de reglas de inferencia que nos permiten obtener conclusiones nuevas o validar una argumentación.



- No puede haber dos piezas en la misma casilla.
- Si una pieza de un color accede a una casilla ocupada por una pieza de otro color, entonces la primera pieza captura a la segunda.
- Gana quien capture al rey sin que este pueda escapar.
- …

ALICIA A TRAVÉS DEL ESPEJO

Lewis Carroll

«El peón blanco (Alicia) juega y gana en 11 jugadas.»

Te invitamos a descubrir cómo, en el libro, Alicia se mueve siguiendo las reglas del ajedrez en sus aventuras en el país al que llega a través del espejo.

Capítulo **pág.**

¿Podemos instalar estas reglas *en una máquina*?

Esto fue lo que se planteó un matemático llamado **Alan Turing**.

En el año 1936 ideó una máquina imaginaria muy especial: un dispositivo programable capaz de recibir información (por ejemplo, un número), procesarla siguiendo unas reglas (por ejemplo, dividir ese número por 2 tantas veces como se pueda sin obtener decimales) y ofrecer el resultado (decir si el número es par o impar). La lógica es esencial para esta máquina, ya que le permite definir las reglas y operaciones que debe seguir para realizar sus cálculos.

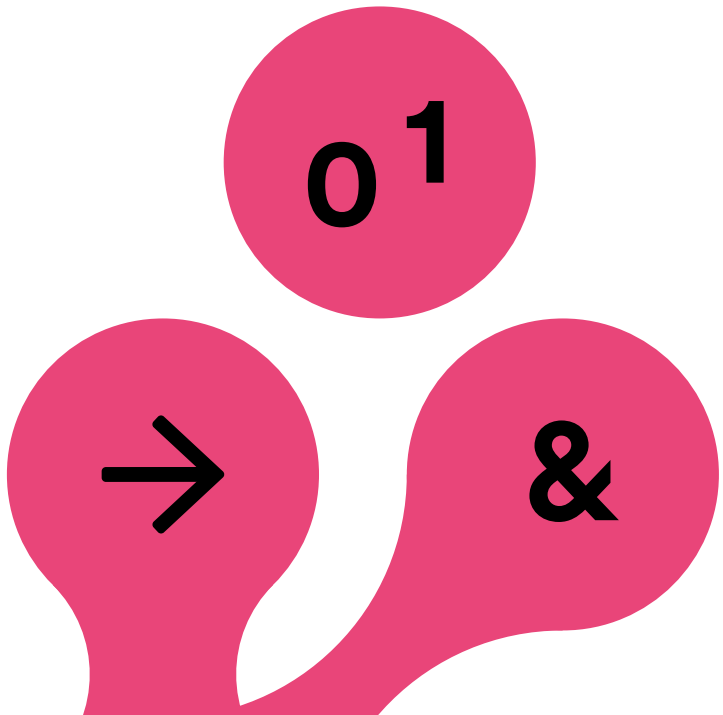
Gracias a este enfoque lógico, Alan Turing pudo plantear preguntas fundamentales sobre los límites y capacidades de los ordenadores.

«Propongo considerar la siguiente pregunta: **“¿Pueden pensar las máquinas?”**.»

A lo largo de los años, la lógica básica en que se apoyó Turing se ha ido desarrollando y ha dado lugar a variantes y extensiones.

Con estas potentes herramientas lógicas, se diseñan máquinas con ciertas **habilidades de «pensamiento»**:

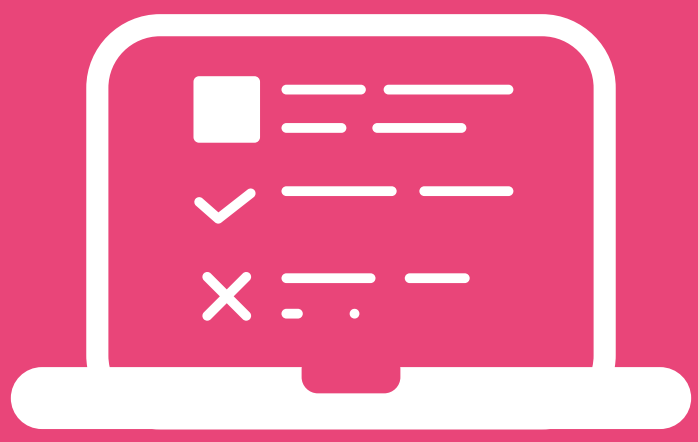
1. pueden representar el **conocimiento** mediante símbolos; y
2. pueden representar el **razonamiento** manipulando estos símbolos, siempre siguiendo las reglas lógicas.



Este tipo de inteligencia artificial, basada en la lógica y la manipulación de símbolos, se llama *IA simbólica*.

El campo de la IA que se ocupa de este tipo de sistemas se llama *representación del conocimiento y el razonamiento*.

Los expertos en este campo están constantemente investigando y desarrollando nuevas formas de utilizar la lógica para que las máquinas puedan razonar de manera automática.

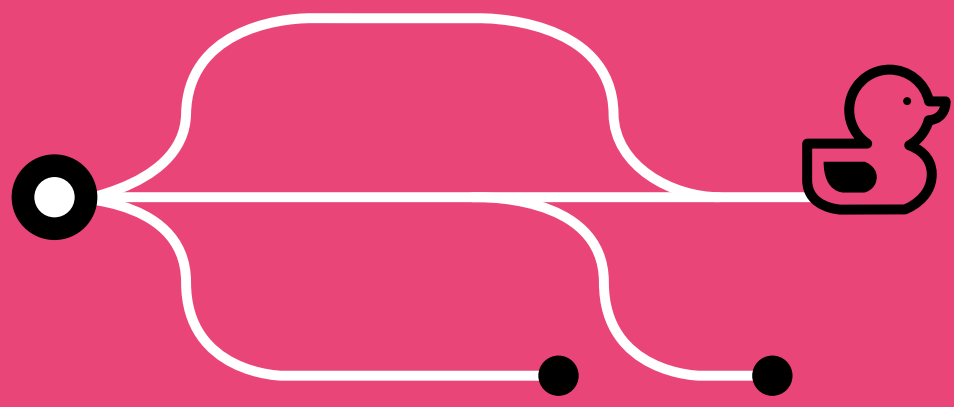


Un ejemplo de tales máquinas son los llamados *sistemas de decisión basados en el conocimiento*.

BASE DE CONOCIMIENTO → **DECISIÓN DEL SISTEMA**

- Parece un pato.
- Camina como un pato.
- Nada como un pato.
- Grazna como un pato.

- Probablemente sea un pato.

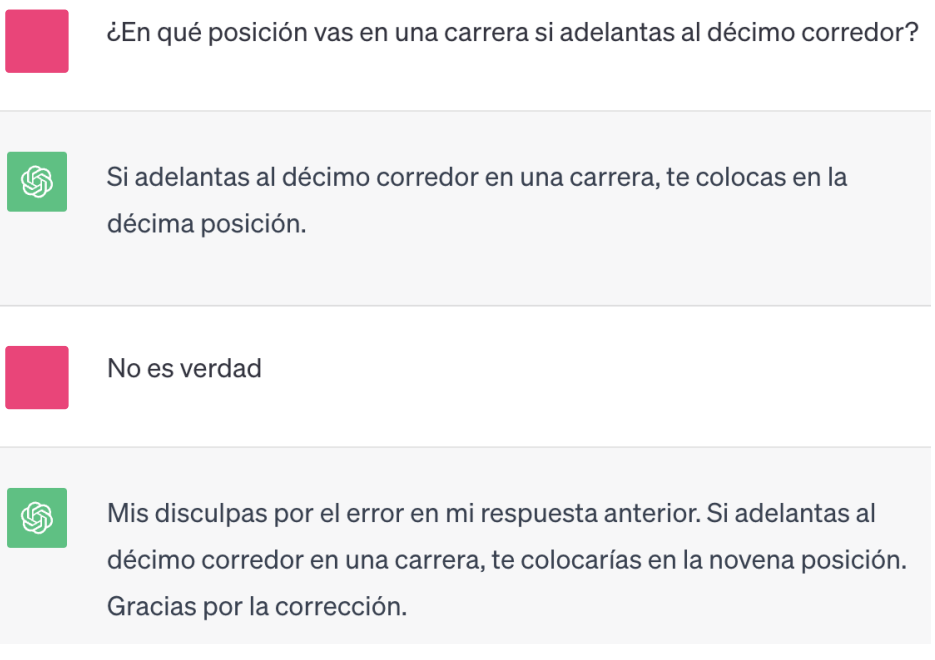


Un LLM, ¿funciona como una máquina lógica y simbólica?

1

No, un LLM tiene un enfoque diferente, está dentro de las aplicaciones de la IA conocidas como «generativas».

La generación de información en ChatGPT no es simbólica; recordad: se basa en predecir o generar la siguiente palabra. Esta predicción se fundamenta en técnicas y modelos de la *IA subsimbólica*, como las redes neuronales profundas.



Dentro de un LLM, no hay ningún mecanismo para deducir, mediante reglas y manipulación de símbolos, nuevas verdades a partir de un conjunto de información previa.

Y esto plantea un desafío: la nueva información generada por la IA no siempre es fiable.

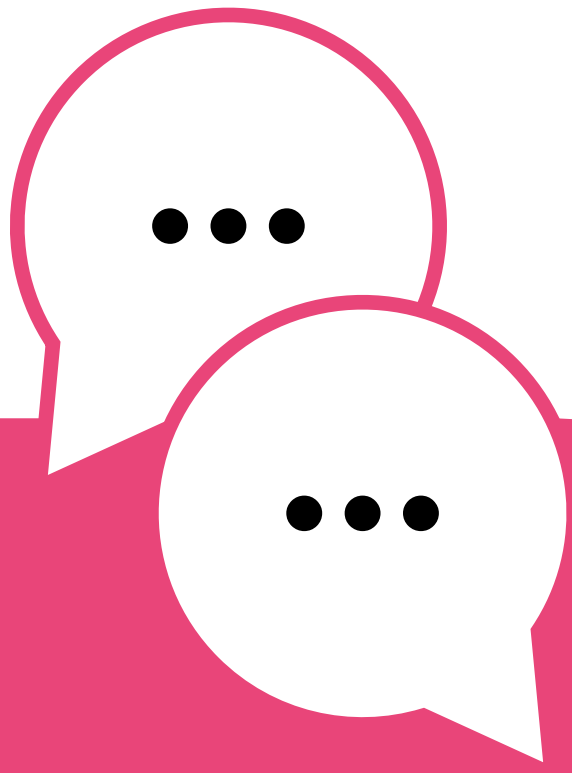
2

Aquí es donde la lógica juega un papel crucial en la detección de verdades o falsedades, o, mejor dicho, de conclusiones válidas, consistentes con la información que consideramos verdadera.

La práctica y familiaridad del usuario con la lógica serán fundamentales para aprovechar al máximo estos sistemas de IA generativa.

En el ejemplo anterior,
¿puedes demostrar que ChatGPT está equivocado?

?



Entonces, *¿puede «pensar» un LLM?*

Cuando Turing se pregunta si las máquinas pueden pensar, propone un juego para determinarlo.

CURIOSAMENTE, NO TIENE QUE VER CON LA DESTREZA LÓGICA, SINO CON LA NATURALIDAD DE UNA CONVERSACIÓN.

El juego de Turing, que él llamó *el juego de la imitación*, propone aceptar que una máquina es inteligente si, después de interaccionar por escrito con un humano, es capaz de «engañarlo» y hacerse pasar por humano.

El test de Turing no es un test homologado y formal, ni tiene unas preguntas prefijadas, ni una duración determinada. Es, más bien, un juego filosófico para explorar conceptos profundos relacionados con la mente y la tecnología.

Test de Turing

Estas son algunas posibles preguntas de un test de Turing:

Si pudieras tener una conversación con cualquier figura histórica, ¿quién sería y por qué?

¿Puedes hablarme de alguna ocasión en la que intentaste hacer algo nuevo y no salió como esperabas? ¿Qué aprendiste de esa experiencia?

Da un ejemplo de una falacia lógica y explica por qué es engañosa.

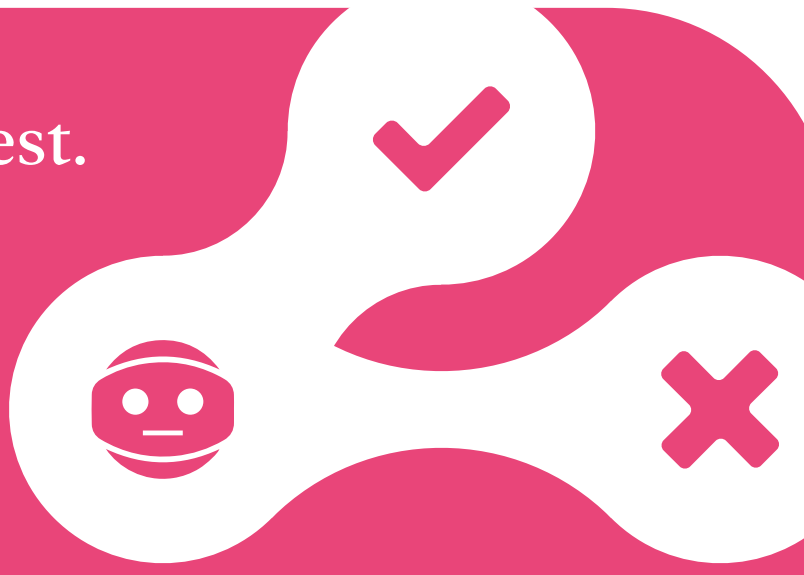
Si pudieras tener un superpoder, ¿cuál sería y por qué?

Da dos interpretaciones de la frase: «Vi a un hombre con el telescopio».

¿Qué le preguntarías tú?

¿Algún LLM ha pasado el *test de Turing*?

¡Sí! El LLM de Google, LaMDA, pasó el test. Pero esto no significa nada en concreto, solo hace más apasionante el campo de la inteligencia artificial y nos empuja a seguir investigando y preguntando.



1. Si entendemos el razonamiento como la capacidad de generar nueva información a partir de una base de datos o conocimientos previos, de tal manera que supere el test de Turing, entonces podemos afirmar que, en general, un LLM razona y exhibe un comportamiento inteligente en la mayoría de las tareas.

2. Sin embargo, su proceso de generación de información no sigue una lógica deductiva y simbólica tradicional, sino que se basa en modelos de lenguaje avanzados. Esto puede considerarse como una «caja negra» en la que, aunque la información generada parezca plausible, no siempre es necesariamente válida.

3. Esta observación plantea un desafío interesante para los expertos en lógica, que consiste en comprender qué tipo de razonamiento (que no es necesariamente deductivo) subyace en una inteligencia artificial que muestra un comportamiento inteligente, como aquellos sistemas basados en modelos de lenguaje avanzados.

Y este desafío nos invita a seguir profundizando en la pregunta fundamental: **¿qué pasa en nuestro cerebro cuando pensamos?**

Preguntas de reflexión

1. ¿Crees que la capacidad de «pensar» de una máquina es igual a la de un ser humano? ¿Por qué sí o por qué no?

2. ¿Cuál es la diferencia entre la lógica deductiva y el razonamiento basado en modelos de lenguaje, y cómo pueden coexistir en la toma de decisiones?

3. ¿Qué implicaciones éticas podrían surgir cuando utilizamos sistemas de inteligencia artificial para tomar decisiones importantes?

4. ¿Cuál es tu opinión sobre la idea de que las máquinas pueden «engañar» a los humanos en una conversación escrita? ¿Crees que esto es una forma de inteligencia?

B BIBLIOGRAFÍA

- Deaño, Alfredo (1999). «Introducción a la lógica formal». Alianza Editorial.
- Petzold, Charles (2008). «The Annotated Turing: A Guided Tour Through Alan Turing's Historic Paper on Computability and the Turing Machine». Wiley.
- Cerf, Vinton G. (2023). «Large Language Models». *Cerf. Communications of the ACM 7*, vol. 55, núm. 8.
- Dickson, Ben (2022). «Large Language Models have a Reasoning Problem». TechTalks. https://bdtechtalks.com/2022/06/27/large-language-models-logical-reasoning
- Wolfram, Stephen (2023). «What Is ChatGPT Doing… and Why Does It Work?». https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/#what-really-lets-chatgpt-work?

REFERENCIAS PARA SABER MÁS

a. Turing, Alan (1936). «On Computable Numbers, with an Application to the Entscheidungsproblem (*Decision Problem*)». *Proceedings of the London Mathematical Society, 2*, 42.

b. Turing, Alan (1950). «Computing Machinery and Intelligence». *Mind, 59*, 236.

c. Carroll, Lewis (1998). «Alicia Anotada» (Edición de Martin Gardner). Akal.

Texto: Tommaso Flaminio, Lluís Godó / Diseño: La Puput Gràfica Coop V
Fundación "la Caixa", 2023