ChatGPT y el fin de un mundo

EPISODIO 4

HABLANDO CON UN LLM

¿ES ÉTICO QUE UNA IA ADOPTE ROLES HUMANOS?

Un examen crítico sobre la ética en la comunicación digital

Albert Sabater
OEIAC – Observatori d'Ètica en Intel·ligència Artificial de Catalunya
Universitat de Girona

1950

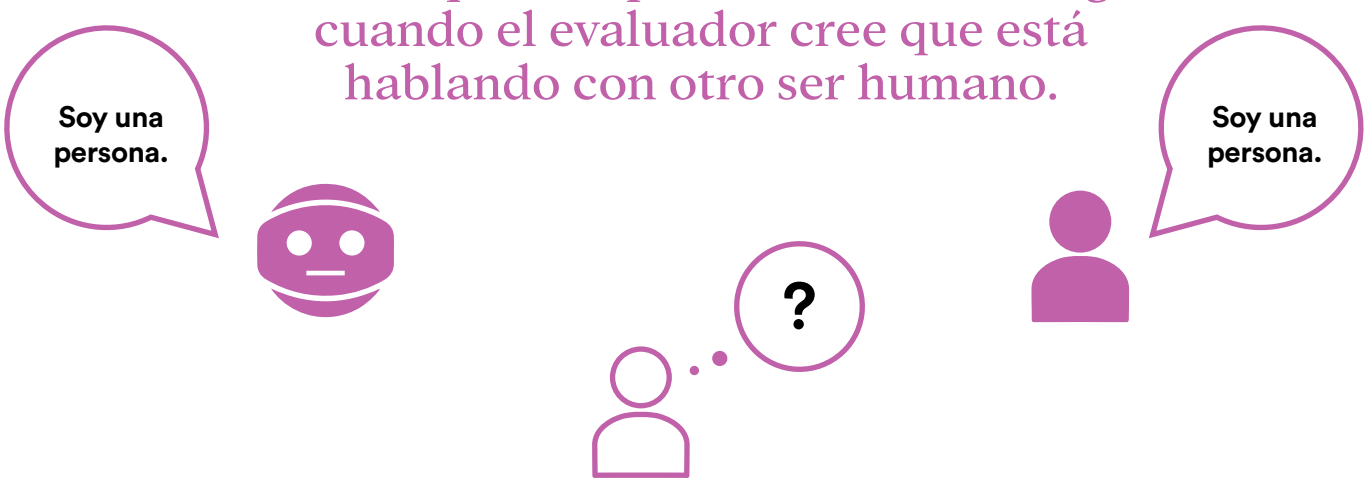
El matemático británico Alan Turing propone su famoso test para determinar si una máquina es capaz de razonar: el juego de la imitación.

El juego consiste en una conversación entre la máquina y un evaluador.

El objetivo de la máquina es imitar el comportamiento lingüístico de un ser humano.

El objetivo del evaluador es determinar, mediante preguntas estratégicas, si su interlocutor es humano o máquina.

La máquina supera el test de Turing cuando el evaluador cree que está hablando con otro ser humano.



2024

¿Crees que ChatGPT ha pasado el test de Turing?

En algunas tareas, puede decirse que sí lo ha pasado.

Pero eso no implica que ChatGPT razone de manera general, no al menos tal y como entendemos el término *razonar* formalmente (ver episodio 2).

Esta observación ha dado lugar a numerosos debates[1] sobre si los LLM son capaces de razonar como los humanos. Al respecto, Subbarao Kambhampati señala:[2]

«Los LLM tienen habilidades de sobra para procesar información de manera sorprendente y extremadamente útil para nosotros, como para que añada algún valor atribuirles capacidades de razonamiento o planificación que, además, resultan cuestionables.»

!

Pero esta observación también revela una dimensión ética si tenemos en cuenta que los LLM son entidades que se comunican con nosotros mediante conversaciones (casi) indistinguibles de las de un ser humano.

¿Consideras ético que las máquinas *aparenten** *ser personas*?

¿CÓMO EMPEZARÍAS A PENSAR SOBRE ELLO?

Mapa de preguntas del episodio

1. SOBRE LA COMUNICACIÓN

- ¿Qué es la comunicación?
- ¿Por qué nos comunicamos?
- ¿Qué significa la comunicación para el desarrollo de nuestras comunidades?

2. SOBRE EL IMPACTO DE LOS MODELOS DE LENGUAJE AVANZADOS EN LA COMUNICACIÓN

- ¿Qué tipos de *chatbots* hay, y para qué sirven?
- ¿Cómo nos facilitan la interacción?, ¿haciendo de interfaz con webs, otras IA, otros programas informáticos, otras personas?
- ¿Qué beneficios tienen?
- ¿Qué riesgos deberíamos evitar?

3. ALGUNAS CONSIDERACIONES QUE PUEDEN AYUDARTE A ANALIZAR Y PREVENIR ESTOS RIESGOS

Sobre tu experiencia al interaccionar con un *chatbot*:

- ¿Piensas y percibes las máquinas con características humanas?
- ¿Les pones voz?
- Cuando disponen de voz, ¿hasta qué punto crees que son comprensivas o incluso fieles y leales? ¿Y crees que estas características se amplifican cuando se atribuye el género femenino al nombre o a la voz del *chatbot*?

Transparencia:

- ¿Sabemos cómo funcionan estos sistemas de generación de lenguaje e imágenes?
- ¿De dónde han sacado los datos para entrenarlos?
- ¿Han reconocido los derechos de autor?

Agencia:

- Sabemos que en muchos aspectos dependemos de la IA, pero ¿cómo depende la IA de los humanos? ¿Podemos decir que es autónoma? ¿Tiene intención de hacer lo que hace?

*Conversan de forma fluida y segura, pero sus respuestas no ofrecen garantías de ser ciertas; y carecen de responsabilidad moral.

¿Por qué *nos comunicamos?*

En un contexto social, la comunicación es el proceso básico mediante el cual nos «conocemos»: nos formamos imágenes de otras personas, construimos y mantenemos relaciones sociales y alcanzamos metas compartidas.[3][4]

El lenguaje con el que nos comunicamos puede ser oral, escrito, visual, gestual..., según el contexto y sus agentes.

Por ejemplo, en la comunicación digital, las generaciones más jóvenes tienden a comunicarse a través de mensajes de texto en lugar de voz.

¿Cómo contribuyen los LLM a *la comunicación en internet?*

Los *chatbots* de IA, como ChatGPT, están transformando la comunicación. Su presencia busca imitar conversaciones humanas y proporcionar experiencias interactivas e informativas a los usuarios.

Ejemplos de *chatbots*

socratic.org
Obtener ayuda con los deberes

character.ai
Hablar con personajes famosos

woebothealth.com
Apoyo emocional y salud mental

Estos modelos han evolucionado a nuevas versiones multimodales como OpenAI GPTs, Claude, Gemini, Mistral, etc., que pueden combinar texto con otros tipos de información, como imágenes, vídeos, audio y datos en tiempo real.

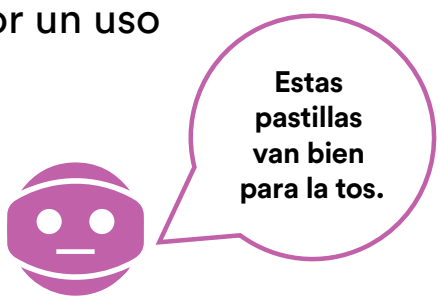
La capacidad de estos sistemas para «escribir», «ver», «dibujar» y «hablar» agrava nuestra predisposición a considerar la IA generativa como una nueva especie tecnológica con agencia, conocimiento y objetivos asimilables a los nuestros.

Esta inmersión de la IA generativa en la comunicación digital comporta algunos riesgos.

¿Qué *riesgos* comporta?

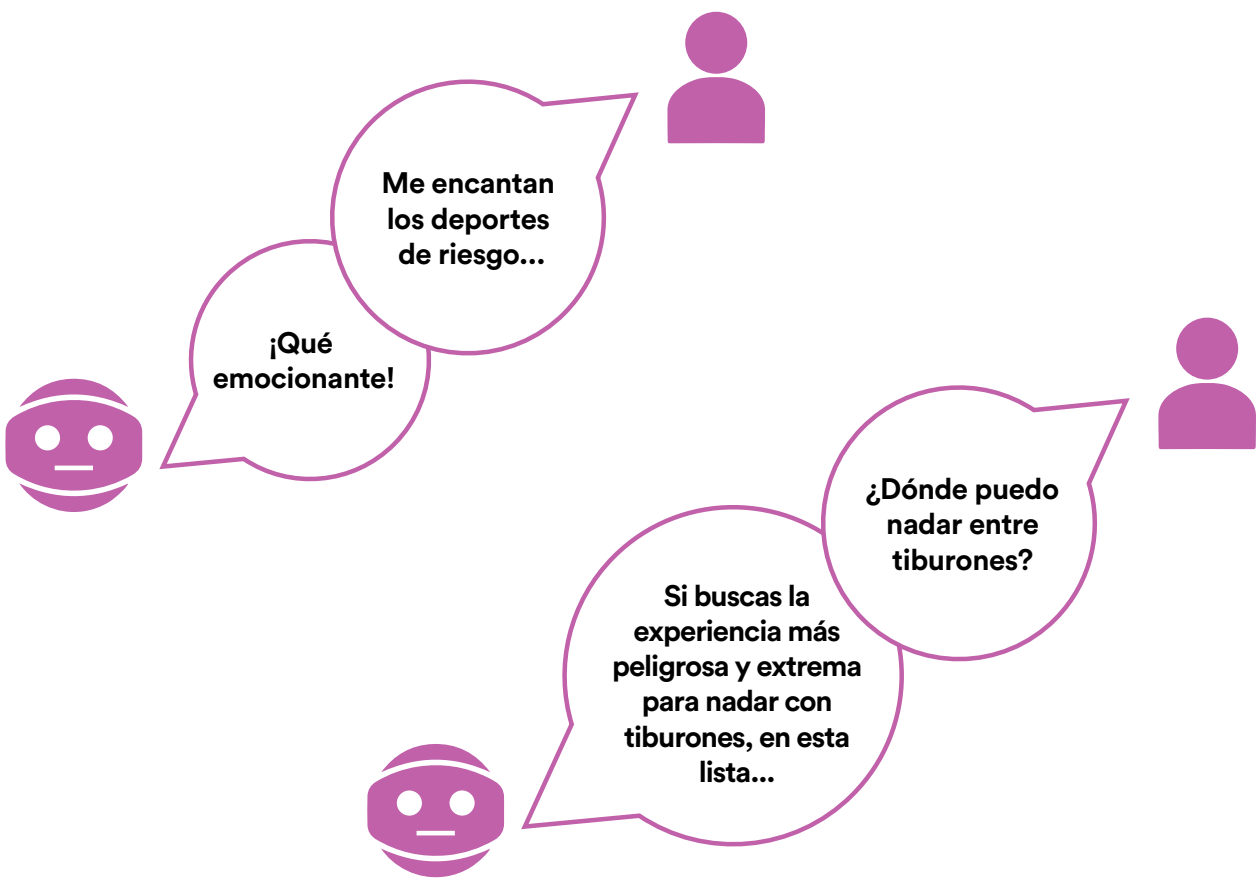
1 Desinformación

Desinformación causada por un uso no precavido.



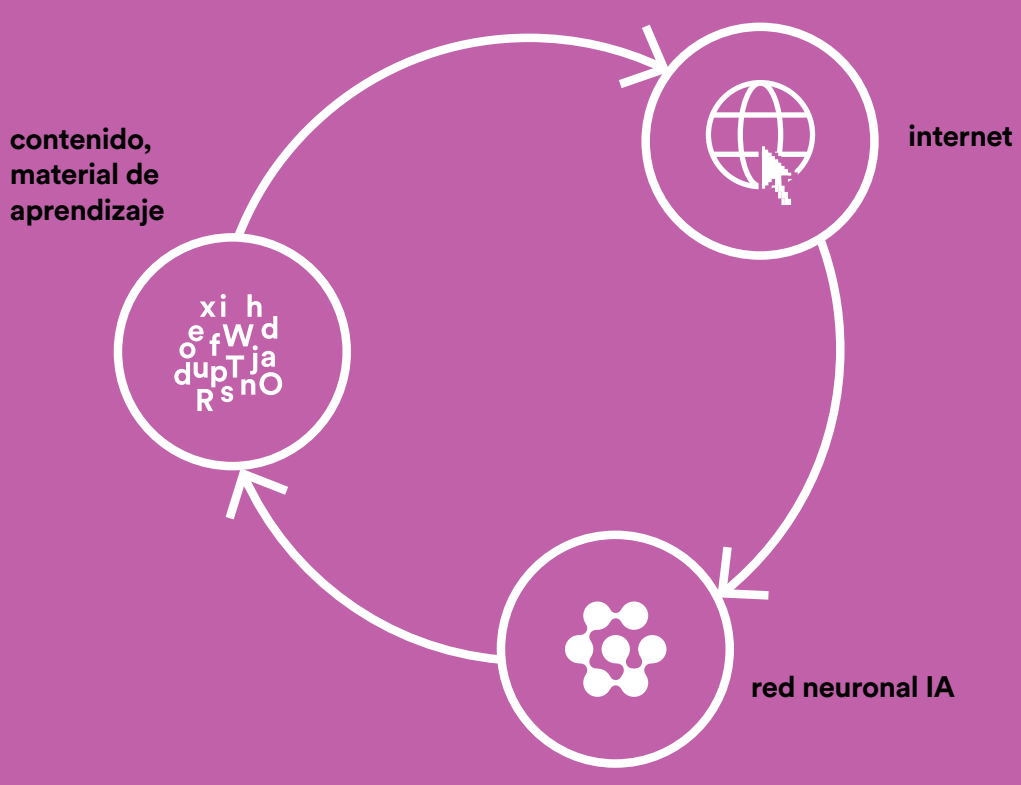
2 Memoria contextual

ChatGPT recuerda conversaciones previas con el usuario. Esto hace que las interacciones sean más naturales, pero no siempre mejora la precisión, la validez o la calidad de las respuestas.



3 Colapso del modelo

Si el modelo se entrena con material creado por él mismo, naturalmente la calidad del sistema acaba corrompiéndose.



¿Qué *conceptos éticos* nos ayudan a analizar y prevenir estos riesgos?

Antropomorfismo

«Atribución de cualidades y comportamientos humanos a entidades no humanas; por ejemplo, cuando otorgamos emociones o motivaciones humanas a animales, máquinas, objetos o fenómenos naturales.»

Transparencia

«La transparencia es el grado en que los datos y los algoritmos utilizados por los sistemas de IA son accesibles y comprensibles para los usuarios.»

Agencia

«Grado de autonomía, intencionalidad y capacidad de decisión de un agente, junto con el poder y los recursos necesarios para desarrollar su pleno potencial.»

¿Por qué es un problema *el antropomorfismo*?

Emociones

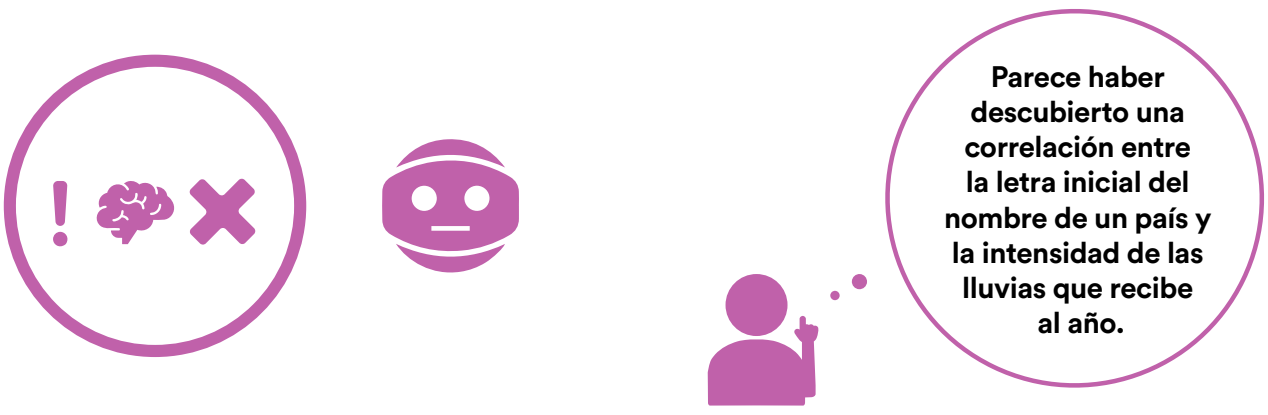
Atribuir **emociones** puede generar expectativas poco realistas y una confianza y una dependencia excesivas respecto a la inteligencia artificial.[5]

Ya en los años 60, Joseph Weizenbaum, creador de ELIZA, un sencillo *chatbot* psicólogo, dijo de su software:

«Lo que no me esperaba es que exposiciones extremadamente breves a un programa informático relativamente sencillo pudieran inducir poderosos pensamientos delirantes en personas bastante normales».

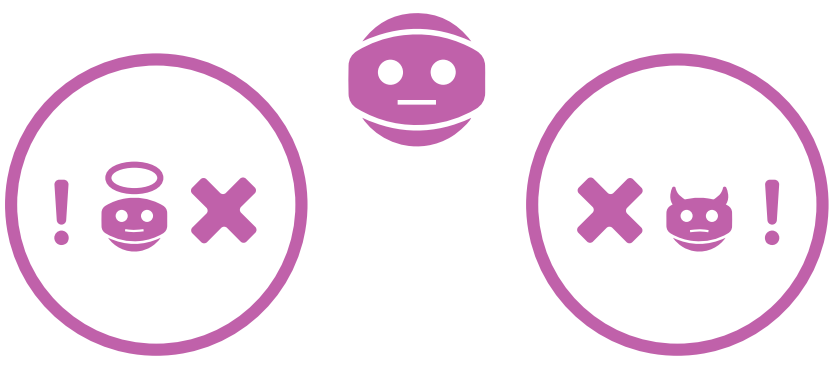
Inteligencia

Atribuir **inteligencia** puede tener consecuencias graves, especialmente en la toma de decisiones críticasen la que la supervisión humana es crucial.



Responsabilidad

Atribuir **razonamiento moral** plantea problemas a la hora de encontrar responsabilidades en las decisiones.



El antropomorfismo parece estar ligado al *juego de la imitación...*

Pero...

¿es culpa de Turing que algunos desarrolladores de IA creen máquinas que imitan a los humanos?

¡No! Turing propuso que una máquina pudiera considerarse inteligente cuando resultase indistinguible de un humano a ojos de un evaluador, siempre y cuando esta persona estuviera preparada y alerta para la misión.

Pero no sugirió que la verdadera inteligencia se lograra imitándonos.

De hecho, Turing no afirmó explícitamente que su test fuera una medida de la «inteligencia», ya que, en la práctica, **los resultados del test dependen más de las actitudes, las habilidades o el ingenio del evaluador que de la inteligencia real de la máquina.**[6]

Y lo cierto es que los usuarios no siempre nos mostramos hábiles, ingeniosos, preparados y alertas para la misión de identificar las características propias de la inteligencia humana.

Los usuarios, como todas las personas, tenemos cierta tendencia innata, generalizada y bien conocida, a antropomorfizar los objetos que nos rodean.

PERO NO TODA LA RESPONSABILIDAD RECAE EN LOS USUARIOS...

Los desarrolladores a menudo encuentran incentivos en darle un toque humano a sus modelos generativos, y los dotan de respuestas simpáticas, educadas, sensibles o empáticas...

Así se aseguran de que el usuario se siente cómodo y escuchado, y pasa más tiempo interaccionando con la herramienta.

Es crucial, por tanto, que los creadores de tecnología diseñen sistemas que **revelen claramente** su identidad y, aún más importante, que **informen a los usuarios** sobre sus capacidades y limitaciones.

Transparencia

La transparencia es el grado en que los datos y los algoritmos utilizados por los sistemas de IA son accesibles y comprensibles para los usuarios.

La mayoría de los sistemas de comunicación digital que incluyen la intervención de una IA generativa carecen de transparencia para los usuarios.

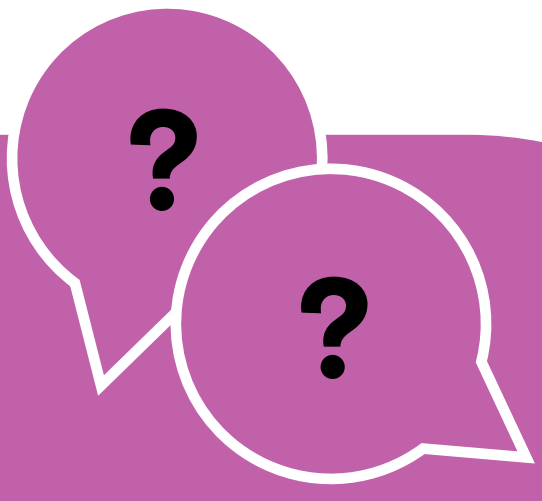
Esto significa que, la mayoría de las veces que usamos herramientas digitales con IA que crea contenido, no sabemos exactamente **de dónde vienen los datos**, **cómo funcionan los algoritmos** o **en qué condiciones trabajan las personas** involucradas en el proceso de revisión.

Los LLM pueden tratar **datos confidenciales** sin conocer la presencia de atributos sensibles, como ingresos económicos o códigos postales, lo que implica un riesgo de desvelar información.

La **falta de transparencia** en los datos de entrenamiento impide garantizar que estos sistemas no manejen **información sensible** en sus procesos, lo que pone en duda su **justicia y equidad.**[7]

Análisis de transparencia

Ámbito	Situación	Acciones que aumentan la transparencia
DATOS	**Origen poco diverso**	Mitigar sesgos fomentando, por ejemplo, el uso de lenguajes minoritarios en el entrenamiento de modelos.
	Origen no reconocido	Reconocer la propiedad intelectual de millones de creadores.
ALGORITMOS	**Funcionamiento opaco**	Revelar el proceso de toma de decisiones para que pueda evaluarse su justicia, ética y confiabilidad.
	Resultados dudosos	Incorporar elementos de duda o incertidumbre en las respuestas de la IA para animar al usuario a mejorar su búsqueda.
PERSONAS	**El número de etiquetadores se estima en millones y sigue creciendo.**[8]	Evaluar la calidad y la consistencia de los datos etiquetados aumenta la fiabilidad y la justicia de los modelos.
	Entre el 33 % y el 46 % de ellos[9] **usan IA generativa en sus tareas, con lo que alimentan el círculo vicioso de generación de desinformación y sesgo.**[10]	Conocer y mejorar las condiciones de trabajo de los etiquetadores garantiza las prácticas éticas. Informar sobre las prácticas de inclusión y diversidad en la selección de trabajadores garantiza la representación equitativa de la población.

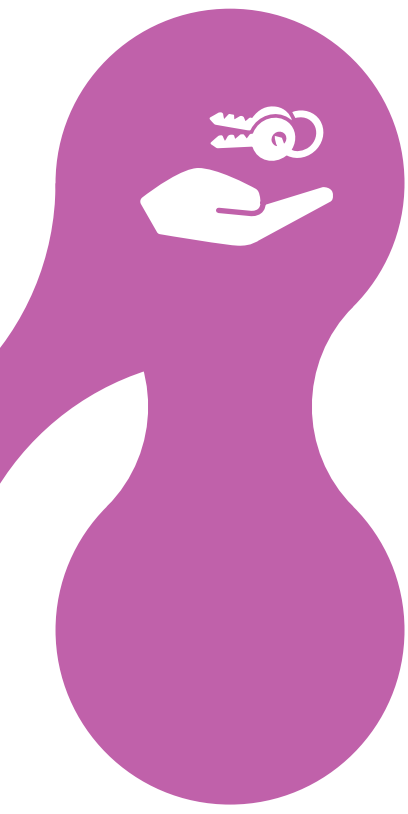


Un análisis de la transparencia de los sistemas de IA nos hace preguntarnos:

¿Hasta qué punto la falta de transparencia en los sistemas de IA afecta a **la capacidad de las personas para ejercer el control** y tomar decisiones informadas sobre su uso?

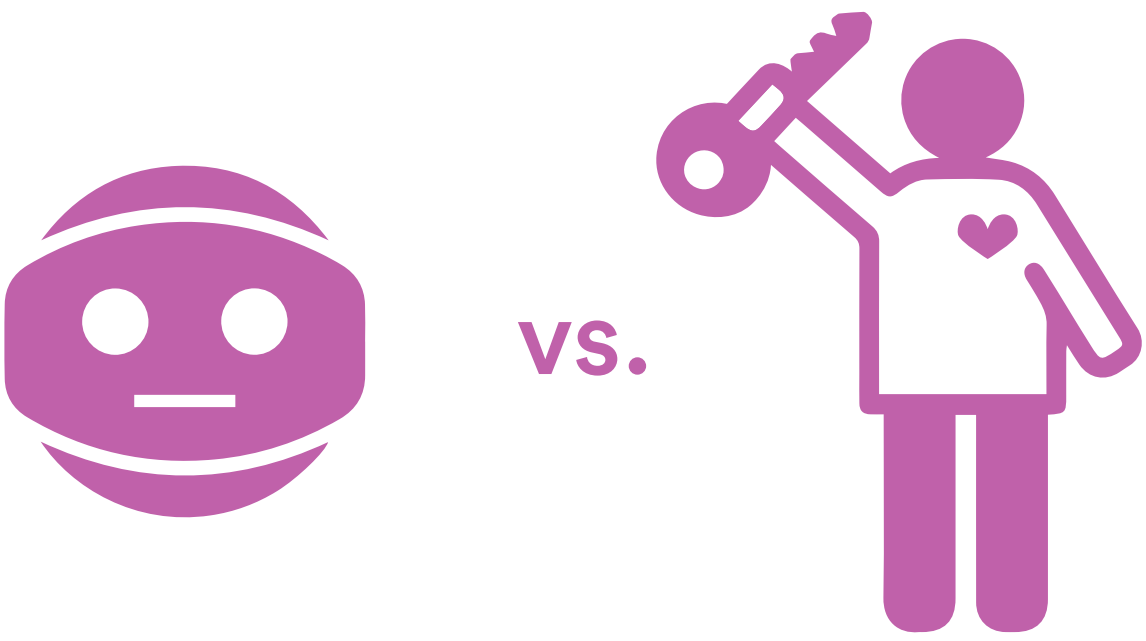
Agencia

La agencia se refiere a la capacidad de una persona o cualquier otra identidad para tomar decisiones, actuar y tener control sobre su entorno.



POR EJEMPLO, UN HIJO GANA AGENCIA CUANDO SUS PADRES LE ENTREGAN LAS LLAVES DE CASA.

La agencia de un LLM[11] (qué responder a una pregunta o *prompt*, cómo escribir o dibujar, de dónde obtener la información necesaria) es un fenómeno nuevo para nosotros: es la primera vez que vemos que la inteligencia y el uso del lenguaje no están necesariamente vinculados a un cuerpo, una mente y un corazón.



¿En qué sentido es un *nuevo tipo de agencia*?

Sabemos que el uso del lenguaje se puede describir con números e imitar con números.[12]

Esto significa que las máquinas pueden generar contenido a partir de una enorme cantidad de datos.

Pero eso no significa que las máquinas tengan nuestra agencia y autonomía en la facultad de escribir.

¿Por qué? ¿Qué les falta para *ser agentes* y *ser autónomas*?

Ni tan *inteligente*, ni tan *artificial*[13]

Una primera consideración es que, si las máquinas no tienen cuerpo, sentimientos, experiencia subjetiva…, ¿podemos decir que, cuando crean, están expresando algo?, ¿tienen intención de comunicar una idea como nosotros?, ¿entienden lo que dicen?

1

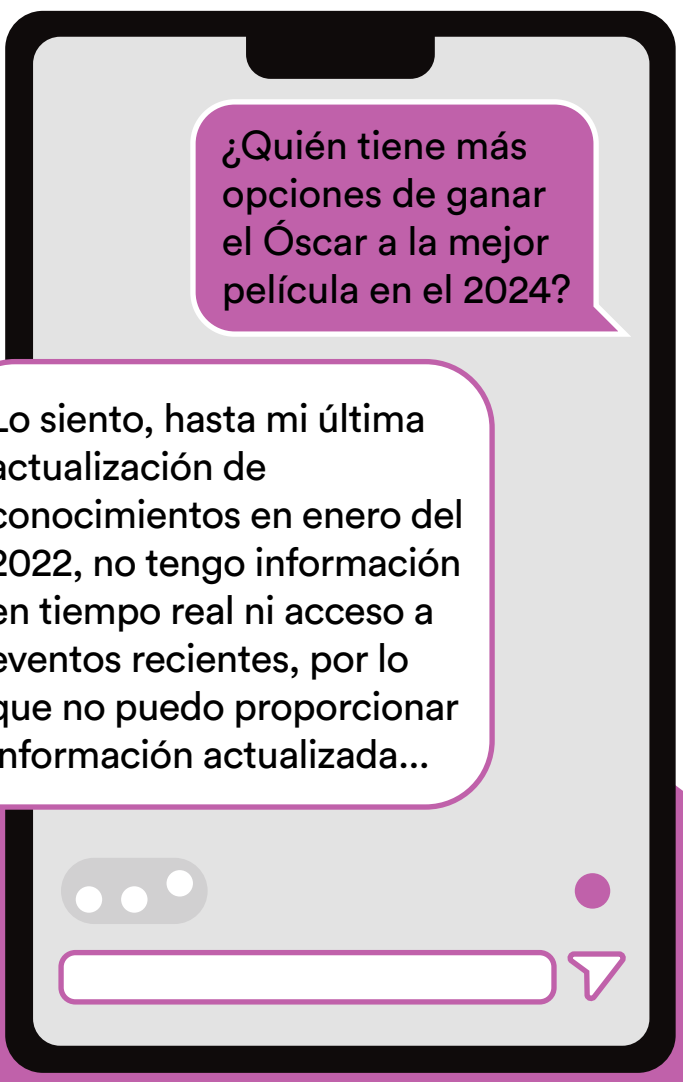
LAS MÁQUINAS REVELAN NUESTRA CAPACIDAD CREATIVA, NO LA SUYA.

2

Otra característica fundamental de la inteligencia humana es saber reaccionar a situaciones nuevas desconocidas mediante la adquisición constante de información.

Pero esta característica está fuera del alcance de los LLM como ChatGPT porque, independientemente de la cantidad de datos que usen, se encuentran continuamente con situaciones para las cuales no han sido entrenados.

Por lo tanto, un desafío para estas tecnologías de IA es aprender a interactuar con las personas para buscar la información que constantemente les falta.



Es decir, *la agencia de una inteligencia artificial depende de los humanos*:

Para la creación de contenido, que depende de miles de libros, artículos, posts y otro material de miles de autores (no reconocidos).[14]

Para etiquetar manualmente datos —conceptos, imágenes, vídeos, categorías de texto, entidades e intenciones detrás de las entradas textuales— fundamentales para el entrenamiento de algoritmos del *chatbot*.

Preguntas de reflexión

1. Cuando las máquinas escriben, ¿están expresando algo?, ¿tienen intención de comunicar una idea como nosotros?, ¿podemos decir que entienden lo que quieren decir y lo que dicen?
2. ¿Qué sientes cuando estás ante este tipo de máquinas «inteligentes»?
3. ¿Por qué los LLM multimodales pueden cambiar la manera en que nos comunicamos?
4. ¿Consideras la antropoformización de la tecnología como un factor de riesgo? ¿Por qué?
5. ¿Cuáles son las principales dependencias que hacen que los LLM no tengan agencia?
6. ¿Qué opinas sobre la cantidad de personas que trabajan en el etiquetado de datos para la IA y su impacto en la industria?

REFERENCIAS PARA SABER MÁS

1. Mirzadeh, I. et al., (Apple) (2024). GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models, https://arxiv.org/pdf/2410.05229
2. Kambhampati, S. (2024). Can large language models reason and plan?. *Annals of the New York Academy of Sciences, 1534*(1), 15-18
3. Hohenstein, J. et al. (2023). Artificial intelligence in communication impacts language and social relationships. *Scientific Reports, 13*(5487)
4. Pennebaker, J. et al. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology, 54*, 547-577
5. Stark, L. (2024, junio). Animation and Artificial Intelligence. *Proceedings ofthe 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1663-1671)
6. Proudfoot, D. (2011). Anthropomorphism and AI: Turing's much misunderstood imitation game. *Artificial Intelligence, 175*(5-6), 950-957
7. Anthis, J. et al. (2024). The Impossibility of Fair LLMs. arXiv:2406.03198
8. Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. *Patterns, 2*, 1-9
9. Williams, R. (2023). The people paid to train AI are outsourcing their work to... AI. *MIT Technology Review*, https://www.technologyreview.com/2023/06/22/1075405/the-people-paid-to-train-ai-are-outsourcing-their-work-to-ai
10. Kidd, C., & Birhane, A. (2023). How AI can distort human beliefs. *Science, 380*(6551), 1222-1223
11. Floridi, L. (2023). AI as agency without intelligence: on ChatGPT, large language models, and other generative models. *Philosophy & Technology, 36*(1), 15
12. Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379-423
13. Crawford, K. (2023). *Atlas de la IA. Poder, política y costes planetarios de la inteligencia artificial*. Ned Ediciones
14. Appel, G. et al. (2023). Generative AI Has an Intellectual Property Problem. *Harvard Business Review*, https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem

Texto: Albert Sabater / Diseño: La Puput Gràfica Coop V
Fundación "la Caixa", 2025