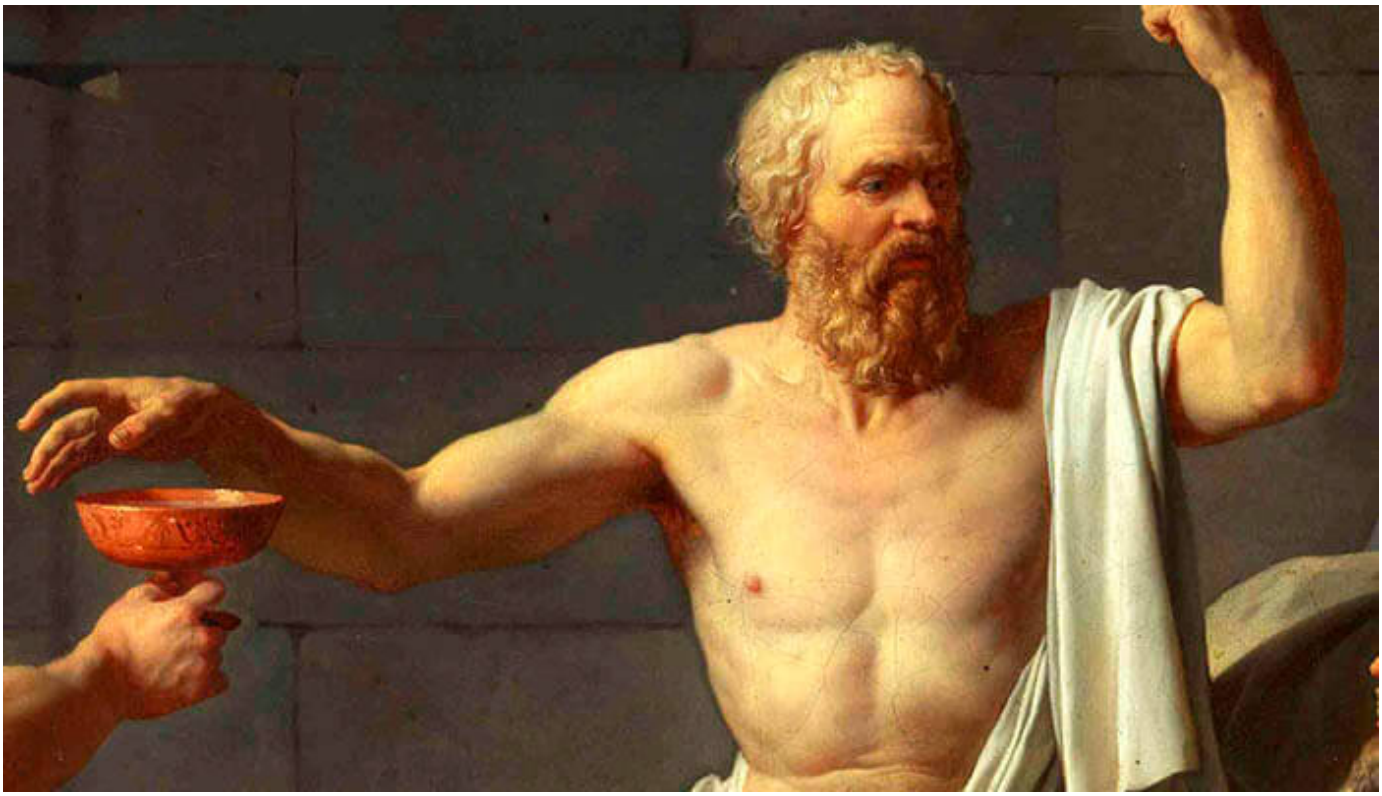

LO QUE SÓCRATES PUEDE ENSEÑARNOS SOBRE LA IA

Carissa Véliz

Si Sócrates fue la persona más sabia de la antigua Grecia, entonces los grandes modelos lingüísticos deben ser los sistemas más estúpidos del mundo moderno. En su *Apología,* Platón cuenta la historia de cómo el amigo de Sócrates, Querefonte, va a visitar el oráculo de Delfos. Querefonte pregunta al oráculo si hay alguien más sabio que Sócrates. La sacerdotisa responde que no, que Sócrates es el más sabio de todos.

En un principio, Sócrates parece desconcertado. ¿Cómo podría él ser el más sabio, cuando había tantas otras personas que eran bien conocidas por su conocimiento y sabiduría, y sin embargo Sócrates afirma que carece de ambos?

Se marca como objetivo resolver el misterio. Va por ahí interrogando a una serie de políticos, poetas y artesanos (como hacen los filósofos). ¿Y qué encuentra? La investigación de Sócrates revela que aquellos que afirman tener conocimiento no saben realmente lo que creen que saben, o saben mucho menos de lo que dicen saber.

Sócrates es el más sabio, entonces, porque es consciente de los límites de su propio conocimiento. No cree saber más de lo que realmente sabe, ni afirma saber más de lo que realmente sabe.

¿Cómo se compara eso con grandes modelos de lenguaje como ChatGPT4?

1 A diferencia de Sócrates, los grandes modelos de lenguaje no saben lo que no saben. Estos sistemas no están diseñados para rastrear la verdad. No se basan en evidencia empírica ni en lógica. Hacen conjeturas estadísticas que muy a menudo son erróneas.

A DIFERENCIA DE SÓCRATES, LOS GRANDES MODELOS DE LENGUAJE NO SABEN LO QUE NO SABEN. ESTOS SISTEMAS NO ESTÁN DISEÑADOS PARA RASTREAR LA VERDAD. NO SE BASAN EN EVIDENCIA EMPÍRICA NI EN LÓGICA. HACEN CONJETURAS ESTADÍSTICAS QUE MUY A MENUDO SON ERRÓNEAS.

Los grandes modelos de lenguaje no informan a los usuarios de que están haciendo conjeturas estadísticas. Presentan conjeturas incorrectas con la misma seguridad con que presentan hechos. Preguntes lo que preguntes te darán una respuesta convincente, y nunca será un “no lo

sé”, aunque debería serlo. Si preguntas a ChatGPT sobre los acontecimientos actuales, te recordará que solo tiene acceso a información hasta septiembre de 2021y que no puede navegar por internet. Para casi cualquier otro tipo de pregunta, se aventurará a dar una respuesta que a menudo mezclará hechos con confabulaciones.

2 El filósofo Harry Frankfurt aportó el célebre argumento de que la estupidez suele ser un discurso persuasivo pero que está desconectado de la preocupación por la verdad. Los grandes modelos de lenguaje son la mayor de las estupideces porque están diseñados para ser plausibles —y por tanto convincentes— sin tener en cuenta la verdad. La estupidez no necesita ser falsa. A veces los estúpidos describen las cosas tal y como son, pero si no buscan la verdad, lo que dicen sigue siendo una tontería.

Y la tontería es peligrosa, advirtió Frankfurt. Esta es una amenaza mayor para la verdad que la mentira. La persona que miente cree saber cuál es la verdad y, por tanto, se preocupa por ella. Puede ser desafiada y responsabilizada; su agenda puede inferirse. El que dice la verdad y el mentiroso juegan en bandos opuestos del mismo juego, como dice Frankfurt. Este último no presta atención al juego. La verdad ni siquiera es enfrentada. Se ignora. Se vuelve irrelevante.

3 La estupidez es más peligrosa cuanto más persuasiva, mientras los grandes modelos de lenguaje son persuasivos por diseño en dos sentidos. En primer lugar, han analizado enormes cantidades de texto, lo que les permite hacer una suposición estadística sobre cuál es la respuesta más apropiada al prompt dado. En otras palabras, imita los patrones que ha captado en los textos que ha recorrido. En segundo lugar, estos sistemas se refinan a través de un proceso de aprendizaje por refuerzo a partir de retroalimentación humana (RLHF). El modelo de recompensa se ha entrenado directamente a partir de la

retroalimentación humana. Los humanos le enseñaron qué tipo de respuestas prefieren. A través de numerosas iteraciones, el sistema aprende a satisfacer las preferencias de los seres humanos, volviéndose cada vez más persuasivo.

Como nos ha enseñado la proliferación de noticias falsas, los seres humanos no siempre preferimos la verdad. Algo falso suele resultar mucho más atractivo que una verdades insulsa. Nos gustan mucho más las buenas historias, las emocionantes, que la verdad. Los grandes modelos de lenguaje son como un mal estudiante, o un mal profesor o un mal periodista que, en lugar de reconocer los límites de su conocimiento, intentan salir del paso engañándote.

4 La *Apología* de Platón sugiere que deberíamos construir la IA para que sea más como Sócrates y menos como estos liantes. No deberíamos esperar que las empresas tecnológicas diseñen éticamente por propia voluntad. Silicon Valley es bien conocido por sus capacidades embaucadoras, y sus empresas pueden sentirse obligadas a mentir para no perder competitividad en su entorno. Que las empresas que trabajan en un entorno corporativo embaucador creen productos embaucadores no debería sorprender a nadie. Una de las cosas que hemos aprendido las últimas dos décadas es que la tecnología necesita tanta regulación como cualquier otra industria, y que ninguna industria puede regularse a sí misma. Regulamos los alimentos, los medicamentos, las telecomunicaciones, las finanzas, el transporte… ¿por qué no debería ser la tecnología la siguiente?

Platón nos deja una última advertencia. Una de las lecciones de su trabajo es tener cuidado con los defectos de la democracia. La democracia ateniense mató a Sócrates. Condenó a su ciudadano más comprometido, a su maestro más valioso, mientras permitió que los sofistas —los embaucadores de aquella época— prosperaran.

Nuestras democracias parecen igualmente vulnerables a los embaucadores. Recientemente los hemos nombrado primeros ministros y presidentes. Y ahora estamos alimentando el poder de los grandes modelos de lenguaje, considerando su uso en todos los ámbitos de la vida, incluso en contextos como el periodismo, la política y la medicina, donde la verdad es vital para la salud de nuestras instituciones. ¿Es eso sabio?

CARISSA VÉLIZ

Carissa Véliz es profesora asociada de la Facultad de Filosofía y del Instituto de Ética de Inteligencia Artificial, así como miembro del Hertford College de la Universidad de Oxford. Véliz ha publicado artículos en *The Guardian*, *The New York Times*, *New Statesman* y *The Independent*. Su trabajo académico se ha publicado en *The Harvard Business Review*, *Nature Electronics*, *Nature Energy* y *The American Journal of Bioethics*, entre otras revistas. Es la editora del *Oxford Handbook of Digital Ethics*. Su último libro es *Profecía. Lecciones sobre el uso y abuso de la predicción, desde los antiguos oráculos hasta la IA* (Penguin Random House).