Un gran poder y una gran responsabilidad

Lo que nos hace especiales a los humanos es nuestra inteligencia. No sabemos muy bien cómo funciona, pero está claro que lo hace, y es lo que nos ha permitido entre otras cosas llegar a la luna. Antes de nosotros, durante eones, las especies habían competido entre sí: si un depredador desarrollaba garras, la presa desarrollaba caparazones; si una especie se volvía venenosa, sus cazadores desarrollaban una tolerancia al veneno. Era una competición sin fin, con muchos perdedores pero ningún ganador.

Hasta que hace unos millones de años, nuestros antepasados lejanos empezaron a desarrollar cerebros más grandes. Cerebros que por otro lado consumían mucha energía, y que, como mucho, servían para comunicarte mejor con otros cerebros, o para hacer herramientas rudimentarias. Pero en cualquier caso, herramientas eran claramente insuficientes para poner nada en la luna. Y sin embargo, es esa inteligencia la que nos ha permitido dominar el mundo.

¿Qué es la inteligencia? A falta de un entendimiento preciso que aún no poseemos, en este texto usaremos este concepto como la capacidad de resolver problemas, especialmente si dichos problemas son de muy diversos tipos. La inteligencia puede, por ejemplo, involucrar comunicación con otros para conseguir objetivos comunes. Y es la inteligencia la que esperamos que nos ayude a desarrollar una teoría unificada de la física, curar enfermedades diversas, o replicar la inteligencia misma en máquinas que llamamos ordenadores.

En este texto, defenderé la tesis de que esta última aplicación puede ser un punto de inflexión en nuestra historia, una tesis extraordinaria que por tanto requiere de evidencia extraordinaria. Considere el lector lo que nos ha permitido hacer la inteligencia hasta ahora, e imagine qué podría pasar si dicha capacidad fuera suficientemente abundante para que pudiera tener un millar de científicos para trabajar día y noche en su problema favorito, sin las limitaciones que nos impone nuestro cuerpo físico. Si la tecnología es la medida del progreso de la humanidad, y la inteligencia humana lo que la produce, ¿qué pasaría si un un producto de dicha tecnología fuese la automatización de la propia inteligencia?

Cómo llegados a este punto entenderá el lector que aunque seamos capaces de entender algunas de las consecuencias de una inteligencia artificial general, usarla sería parecido a ver magia. No podremos seguir sino con mucho esfuerzo sus razonamientos, aunque tan solo sea por la velocidad a la que operan los ordenadores, muy superior a la que podemos razonar nosotros. Por tanto quizá sea apropiado considerar a dichas inteligencias artificiales como pequeños genios a los que les podemos pedir cosas que cumplen.

Ese es un gran poder ¿pero estamos seguros de que nos darán lo que queremos realmente? La literatura está llena de cuentos e historias en los que pedir deseos no acaba bien, empezando por la leyenda del rey Midas. Esto ilustra que algunos deseos pueden tener consecuencias insospechadas si se persiguen con ahínco. ¿Pero donde parar?

El problema reside fundamentalmente en que no sabemos bien siempre qué queremos, o como comunicarlo. Por ejemplo, a algunos nos gustaría estar más en forma, pero cuando es el momento de hacer ejercicio preferimos quedarnos viendo la televisión, o nos gustaría ahorrar más pero no paramos de gastar el dinero que ganamos. Por suerte o por desgracia los seres humanos somos demasiado complicados como para ser descritos unívocamente por un conjunto finito de preferencias. Y todo esto no hace más que complicarse cuando tenemos en cuenta las preferencias de muchas personas, y sus interacciones sociales.

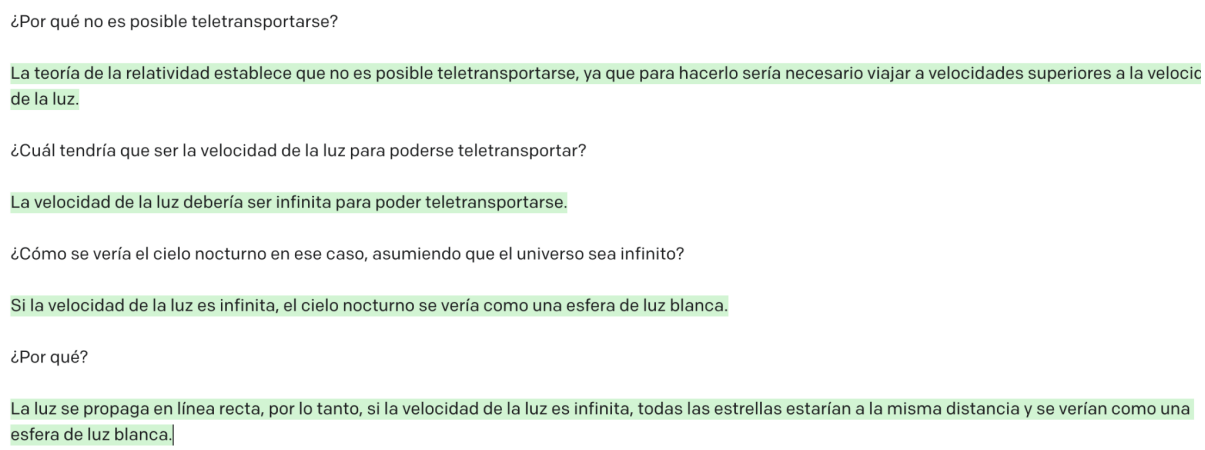
El problema de cómo asegurarnos que una inteligencia artificial persiga nuestros objetivos a pesar de todos estos obstáculos se conoce como el problema del control:

¿Cómo podemos controlar una inteligencia artificial general para que persiga nuestros objetivos de una manera segura?

Hay algunas intuiciones en la comunidad científica sobre cómo abordar este problema. Por ejemplo, el profesor Stuart Russell de la universidad de Berkeley propone que la inteligencia artificial tenga incertidumbre sobre qué queremos y que parte de su objetivo sea aprender qué queremos, no solo de lo que decimos sino también de nuestros actos. Dicho de otra forma, que la propia inteligencia infiera qué queremos a partir de nuestros actos.

Sin embargo, desgraciadamente como sociedad dedicamos muchos más recursos a conseguir sistemas más y más avanzados, y poco a asegurarnos de su seguridad, y de que hagan lo que queremos, con técnicas similares a esta. Es mucho más sexy demostrar las cosas que es capaz de hacer el ultimo sistema disponible, que por ejemplo entender qué es exactamente lo que ha aprendido el modelo mirando dentro.

Por supuesto, la parte técnica del problema del control no es la única. Una tecnología así tiene profundas implicaciones sociales, y por ello la sociedad debería anticiparse a cómo queremos que sea un futuro con esta tecnología a nuestra disposición. ¿Estamos seguros de que siempre habrá trabajo para la mayoría de las personas? ¿Quién debería decidir qué problemas priorizar? ¿Podría llegar a usarse la inteligencia artificial general en conflictos entre países? ¿Qué valores deberían respetar las inteligencias artificiales?



Ejemplo de una conversación con GPT-3, un modelo de inteligencia artificial artificial, desarrollado por OpenAI. La tercera respuesta requirió 3 intentos para acertar.

La inteligencia artificial general puede ser quizá la herramienta más poderosa que llegue a inventar la humanidad, y no parece que esté tan lejos. Los sistemas que tenemos hoy en día son capaces de explicar chistes, resolver algunos problemas de matemáticas de secundaria o pintar imágenes De hecho, hace apenas unas semanas una empresa subsidiaria de Google hizo público que habían desarrollado un sistema que era capaz de hablar, explicar imágenes, jugar a videojuegos y usar brazos robóticos, lo que apunta a indicios de generalidad.

“Osos de peluche trabajando en investigación en inteligencia artificial en la luna en los años 80”. Pintado por DALL-E 2, también desarrollado por OpenAI.

Por supuesto, nada de esto asegura que seremos capaces de conseguir inteligencia artificial pronto, pero no deberíamos confiarnos en que estamos muy lejos para trabajar en su seguridad. Es imperativo que cuando llegue ese día, estemos preparados y hayamos desarrollado las técnicas necesarias para que esta inteligencia sea beneficiosa, tanto desde un punto de vista técnico como social. Si es así, nuestro futuro como sociedad y como especie será brillante, y nosotros somos los responsables de asegurarnos de ello.

Lecturas adicionales

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Christian, B. (2021). *The alignment problem: How can machines learn human values?*. Atlantic Books.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.