LITORALPRESS

Fecha: 04-11-2024
Medio: La Tercera
Supl. : La Tercera
Tipo: Noticia general
Título: Qué es la superinteligencia artificial y por qué se teme que pueda destruir la humanidad

Pág. : 26
Cm2: 760,4

Tiraje: 78.224
Lectoría: 253.149
Favorabilidad: No Definida

► Se trata de software que pueden abordar una gama mucho más amplia de tareas, incluidas cosas como el aprendizaje de nuevas habilidades.

Qué es la superinteligencia artificial y por qué se teme que pueda destruir la humanidad

Según los científicos, la superinteligencia puede estar a sólo "unos pocos miles de días" de distancia.

Flora Salim (The Conversation)*

En 2014, el filósofo británico Nick Bostrom publicó un libro sobre el futuro de la inteligencia artificial (IA) con el ominoso título Superinteligencia: caminos, peligros, estrategias. El libro tuvo una gran influencia en la promoción de la idea de que los sistemas avanzados de IA –"superinteligencias" más capaces que los humanos– podrían algún día apoderarse del mundo y destruir a la humanidad.

Una década después, el jefe de OpenAI, Sam Altman, afirma que la superinteligencia puede estar a sólo "unos pocos miles de días" de distancia. Hace un año, el cofundador de OpenAI de Altman, Ilya Sutskever, creó un equipo dentro de la empresa para centrarse en la "superinteligencia segura", pero él y su equipo han recaudado ahora mil millones de dólares para crear una empresa propia que persiga este objetivo.

¿De qué están hablando exactamente? En términos generales, la superinteligencia es cualquier cosa más inteligente que los humanos. Pero desentrañar lo que eso podría significar en la práctica puede resultar un poco complicado.

Diferentes niveles

En mi opinión, la forma más útil de pensar en los diferentes niveles y tipos de inteligencia en la IA fue desarrollada por la científica informática estadounidense Meredith Ringel Morris y sus colegas de Google.

Su marco de trabajo enumera seis niveles de desempeño de la IA: sin IA, emergente, competente, experta, virtuosa y sobrehumana. También hace una distinción importante entre sistemas limitados, que pueden llevar a cabo una gama pequeña de tareas, y sistemas más generales.

Un sistema estrecho, sin IA, es algo así como una calculadora: lleva a cabo diversas tareas matemáticas según un conjunto de reglas programadas explícitamente.

Ya existen muchos sistemas de IA de bajo nivel que han tenido mucho éxito. Morris cita como ejemplo de un sistema de IA de bajo nivel de virtuosismo el programa de ajedrez Deep Blue que derrotó al campeón mundial Garry Kasparov en 1997.

Algunos sistemas estrechos incluso tienen capacidades sobrehumanas. Un ejemplo es Alphafold , que utiliza el aprendizaje automático para predecir la estructura de las moléculas de proteínas y cuyos creadores ganaron el Premio Nobel de Química este año.

Sistemas generales

Se trata de software que pueden abordar una gama mucho más amplia de tareas, incluidas cosas como el aprendizaje de nuevas habilidades.

Un sistema general sin IA podría ser algo así como el Mechanical Turk de Amazon: puede hacer una amplia variedad de cosas, pero las hace preguntándole a personas reales.

En general, los sistemas de IA general son mucho menos avanzados que sus primos más pequeños. Según Morris, los modelos de lenguaje de última generación detrás de los chatbots como ChatGPT son IA general, pero hasta ahora están en el nivel "emergente" (lo que significa que son "iguales o algo mejores que un humano no capacitado"), y aún no han alcanzado el nivel "competente" (tan bueno como el 50% de los adultos capacitados).

Así que, según este cálculo, todavía estamos a cierta distancia de la superinteligencia general.

¿Qué tan inteligente es la IA en estos momentos?

Como señala Morris, determinar con precisión dónde se encuentra un sistema determinado dependería de disponer de pruebas o puntos de referencia fiables.

Dependiendo de nuestros parámetros de referencia, un sistema de generación de imágenes como DALL-E podría estar en un nivel virtuoso (porque puede producir imágenes que el 99% de los humanos no podrían dibujar o pintar), o podría estar emergiendo (porque produce errores que ningún humano produciría, como manos mutantes y objetos imposibles).

SIGUE ▶▶

LITORALPRESS

Fecha: 04-11-2024
Medio: La Tercera
Supl. : La Tercera
Tipo: Noticia general
Título: Qué es la superinteligencia artificial y por qué se teme que pueda destruir la humanidad

Pág. : 27
Cm2: 785,4

Tiraje: 78.224
Lectoría: 253.149
Favorabilidad: No Definida

SIGUE ▶▶

► Al menos a corto plazo, no tenemos por qué preocuparnos de que una IA superinteligente se apodere del mundo.

Existe un debate importante incluso sobre las capacidades de los sistemas actuales. Un destacado artículo de 2023 afirmaba que el GPT-4 mostraba "destellos de inteligencia artificial general".

OpenAI afirma que su último modelo de lenguaje, o1 , puede "realizar razonamientos complejos" y "rivaliza con el desempeño de los expertos humanos" en muchos puntos de referencia.

Sin embargo, un artículo reciente de investigadores de Apple descubrió que o1 y muchos otros modelos de lenguaje tienen problemas significativos para resolver problemas genuinos de razonamiento matemático. Sus experimentos muestran que los resultados de estos modelos parecen asemejarse a una búsqueda sofisticada de patrones en lugar de a un verdadero razonamiento avanzado. Esto indica que la superinteligencia no es tan inminente como muchos han sugerido.

Más inteligente

Algunas personas creen que el rápido ritmo de progreso de la IA de los últimos años continuará o incluso se acelerará. Las empresas tecnológicas están invirtiendo cientos de miles de millones de dólares en hardware y capacidades de IA, por lo que esto no parece imposible.

Si esto sucede, es posible que veamos una superinteligencia general dentro de los "pocos miles de días" propuestos por Sam Altman (eso es una década o más en términos menos científicos). Sutskever y su equipo mencionaron un período de tiempo similar en su artículo sobre la superalineación.

Muchos de los éxitos recientes en el campo de la IA se han logrado gracias a la aplicación de una técnica llamada "aprendizaje profundo", que, en términos simples, encuentra patrones asociativos en colecciones gigantescas de datos. De hecho, el Premio Nobel de Física de este año ha sido otorgado a John Hopfield y también al "padrino de la IA", Geoffrey Hinton, por su invención de las redes de Hopfield y la máquina de Boltzmann, que son la base de muchos modelos de aprendizaje profundo potentes que se utilizan en la actualidad.

Los sistemas generales como ChatGPT se han basado en datos generados por humanos, muchos de ellos en forma de textos extraídos de libros y sitios web. Las mejoras en sus capacidades se han debido en gran medida a un aumento de la escala de los sistemas y de la cantidad de datos con los que se entrenan.

Sin embargo, es posible que no haya suficientes datos generados por humanos para llevar este proceso mucho más allá (aunque los esfuerzos por utilizar los datos de manera más eficiente, generar datos sintéticos y mejorar la transferencia de habilidades entre diferentes dominios pueden generar mejoras). Incluso si hubiera suficientes datos, algunos investigadores dicen que los modelos de lenguaje como ChatGPT son fundamentalmente incapaces de alcanzar lo que Morris llamaría competencia general.

Un artículo reciente ha sugerido que una característica esencial de la superinteligencia sería su carácter abierto , al menos desde una perspectiva humana. Tendría que ser capaz de generar continuamente resultados que un observador humano consideraría novedosos y de los que podría aprender.

Los modelos básicos existentes no se entrenan de forma abierta, y los sistemas abiertos existentes son bastante limitados. Este artículo también destaca cómo la novedad o la capacidad de aprendizaje por sí solas no son suficientes. Se necesita un nuevo tipo de modelo básico abierto para lograr la superinteligencia.

Cuales son los riesgos

¿Qué implica todo esto para los riesgos de la IA? Al menos a corto plazo, no tenemos por qué preocuparnos de que una IA superinteligente se apodere del mundo.

Pero eso no quiere decir que la IA no presente riesgos. Una vez más, Morris y compañía han pensado en esto: a medida que los sistemas de IA adquieren mayor capacidad, también pueden ganar mayor autonomía. Diferentes niveles de capacidad y autonomía presentan diferentes riesgos.

Por ejemplo, cuando los sistemas de IA tienen poca autonomía y las personas los usan como una especie de consultores (cuando le pedimos a ChatGPT que resuma documentos, por ejemplo, o dejamos que el algoritmo de YouTube dé forma a nuestros hábitos de visualización), podríamos enfrentar el riesgo de confiar demasiado en ellos o depender demasiado de ellos .

Mientras tanto, Morris señala otros riesgos a los que hay que prestar atención a medida que los sistemas de IA se vuelven más capaces, que van desde personas que forman relaciones parasociales con sistemas de IA hasta desplazamientos masivos de puestos de trabajo y aburrimiento en toda la sociedad.

Qué sigue

Supongamos que algún día contamos con agentes de inteligencia artificial superinteligentes y autónomos. ¿Correremos el riesgo de que concentren el poder o actúen en contra de los intereses humanos?

No necesariamente. La autonomía y el control pueden ir de la mano. Un sistema puede estar altamente automatizado y, aun así, ofrecer un alto nivel de control humano. Al igual que muchos en la comunidad de investigación de IA, creo que es posible lograr una superinteligencia segura. Sin embargo, su creación será una tarea compleja y multidisciplinaria, y los investigadores tendrán que recorrer caminos inexplorados para lograrlo.●

Flora Salim, profesora de la Facultad de Ciencias Informáticas e Ingeniería, cátedra inaugural de Cisco de Transporte Digital e IA, UNSW Sydney