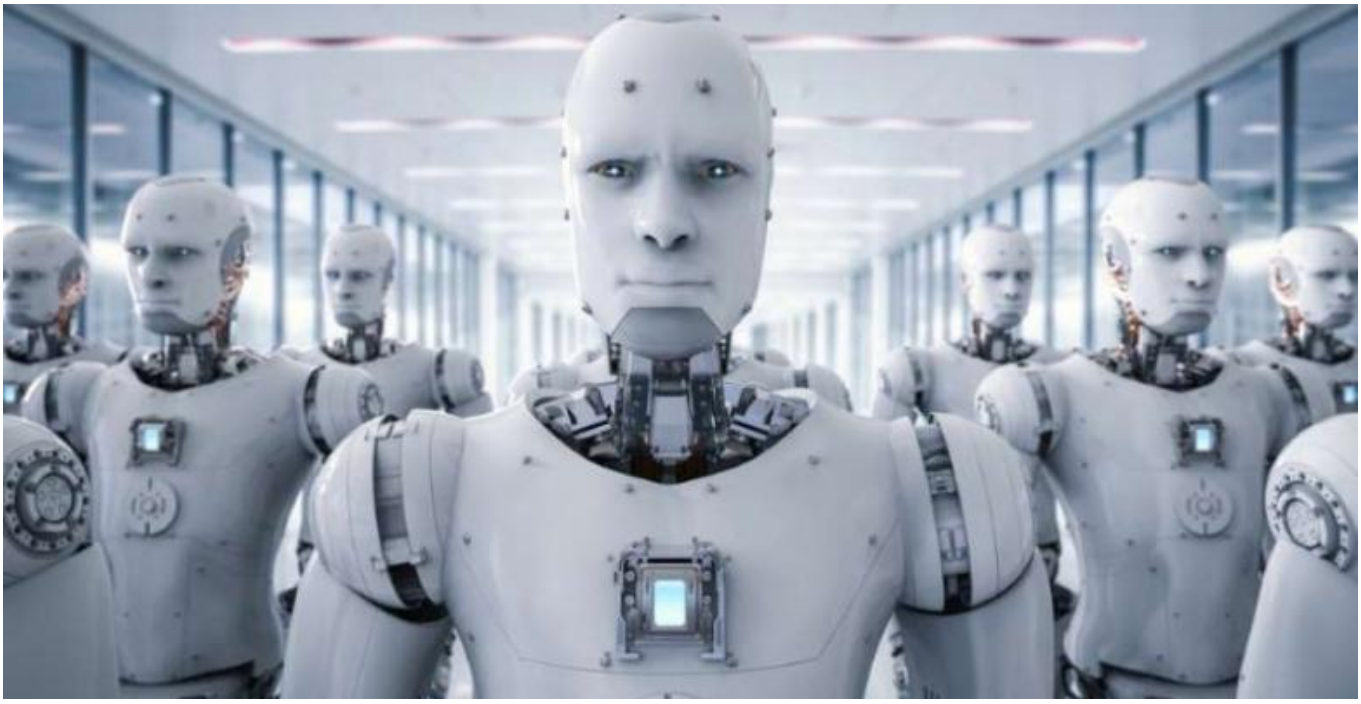
Cómo la inteligencia artificial podría destruirnos por accidente

01/11/2019



Pero según un nuevo libro, lo que debe preocuparnos no es que los robots tomen conciencia de sí mismos y se alcen contra sus amos humanos, sino que las máquinas se vuelvan tan buenas en la consecución de los objetivos que les fijamos, que terminemos siendo aniquilados inadvertidamente al establecerles tareas equivocadas.

Stuart Russell, profesor en la Universidad de California en Berkeley, es el autor de Human Compatible: AI and the Problem of Control ("Compatible con humanos: la IA y el problema del control") y un experto en los avances que el aprendizaje automático ha hecho posibles.

"El meme de Hollywood siempre consiste en la máquina que espontáneamente toma conciencia de sí misma y luego decide que odia a los seres humanos y quiere matarnos a todos", dijo a la BBC.

Pero los robots no tienen sentimientos humanos, por lo que "es completamente equivocado preocuparse por eso".

"No es realmente la conciencia maligna, sino su capacidad la que tiene que preocuparnos, solo su capacidad de alcanzar un objetivo mal especificado por nosotros".

"Demasiado competente"

En una entrevista con el programa Today de la BBC, el experto dio un ejemplo hipotético de la amenaza real que, en su opinión, la IA podría representar.

Imagina que tenemos un poderoso sistema de IA que es capaz de controlar el clima del planeta y que queremos usarlo para devolver los niveles de CO2 en nuestra atmósfera a la época preindustrial.

"El sistema descubre que la forma más fácil de hacerlo es deshacerse de todos los seres humanos, porque ellos son los que están produciendo todo este dióxido de carbono en primer lugar", dijo Russell.

"Y podrías decir, bueno, puedes hacer lo que quieras, pero no puedes deshacerte de los seres humanos. Entonces ¿qué hace el sistema? Simplemente nos convence de tener menos hijos hasta que no queden seres humanos".

El ejemplo sirve para resaltar los riesgos asociados a que la inteligencia artificial actúe bajo instrucciones en las que los humanos no hemos pensado.

Superinteligencia

La mayoría de los sistemas actuales de IA tienen aplicaciones "débiles", diseñadas específicamente para abordar un problema bien especificado en un área, según el Centro para el Estudio del Riesgo Existencial, de la Universidad de Cambridge, en Reino Unido.

Un momento importante para este campo llegó en 1997, cuando la computadora Deep Blue derrotó al campeón mundial de ajedrez, Garry Kasparov, en un torneo de seis partidas.

Pero a pesar de la hazaña, Deep Blue fue diseñado por humanos específicamente para jugar al ajedrez y no podría con un simple juego de damas.

Ese no es el caso de los avances posteriores en inteligencia artificial. El software AlphaGo Zero, por ejemplo, alcanzó un nivel de rendimiento sobrehumano después de solo tres días de jugar Go contra sí mismo.

Usando el aprendizaje profundo, un método de aprendizaje automático que emplea redes neuronales artificiales, AlphaGo Zero requirió mucha menos programación humana y resultó ser un muy buen jugador de Go, ajedrez y sh?gi.

Fue completamente autodidacta, de una manera, tal vez, alarmante.

"A medida que un sistema de inteligencia artificial se vuelva más poderoso y más general, podría volverse súper inteligente, superior al rendimiento humano en muchos o casi todos los dominios", dice el Centro de Riesgo Existencial.

Y es por eso que, según Russell, los humanos necesitamos retomar el control.

"No sabemos lo que queremos"

Según Russell, dar a la inteligencia artificial objetivos más definidos no es la solución para este dilema, porque los humanos mismos no estamos seguros de cuáles son esas metas.

"No sabemos que algo no nos gusta hasta que sucede", dice.

"Deberíamos cambiar toda la base sobre la cual construimos sistemas de IA", dice, alejándose de la noción de dar a los robots objetivos fijos.

"En cambio, el sistema tiene que saber que desconoce cuál es el objetivo".

"Y una vez que tienes sistemas que funcionan de esa manera, realmente serán diferentes a los seres humanos. Comenzarán a pedir permiso antes de hacer las cosas, porque no estarán seguros de si eso es lo que quieres".

En especial, dice el profesor Russell, estarían "felices de que los apaguen porque querrán evitar hacer cosas que no te vayan a gustar".

El genio de la lámpara

"La forma en que construimos la IA es un poco como la forma en que pensamos en un genio dentro de una lámpara. Si frotas la lámpara, sale el genio y dices: 'Me gustaría que esto sucediera'", dijo Russell.

"Y, si el sistema de IA es lo suficientemente potente, hará exactamente lo que pides y obtendrás exactamente lo que pides".

"Ahora, el problema con los genios en las lámparas es que el tercer deseo es siempre: 'Por favor, deshaga los dos primeros deseos porque no pudimos especificar los objetivos correctamente".

"Entonces, una máquina que persigue un objetivo que no es el correcto se convierte, en efecto, en un enemigo de la raza humana, un enemigo que es mucho más poderoso que nosotros".