Cumplir con la misión a cualquier precio

NAIEF YEHYA

Un dron militar, equipado con inteligencia artificial (IA) y entrenado con *software* similar al que se usa para juegos de mesa y de video, tiene por objetivo en una simulación destruir plataformas de lanzamiento de misiles tierra-aire. Debe identificar blancos y solicitar autorización al operador para proceder a destruirlos. Cada vez que cumple con su objetivo gana puntos. Cuando el operador le indica que no debe destruir un blanco que ha identificado, el dron evalúa el mejor curso de acción y decide matar al operador que limita su capacidad de acción para obtener mayor puntaje.

Y eso hace.

Tucker Hamilton, el jefe de pruebas de inteligencia artificial (IA) y operaciones de la Fuerza Aérea estadounidense describió en una conferencia esta simulación con un dron. Sin embargo, poco después la portavoz de la Fuerza Aérea declaró que dicha simulación nunca tuvo lugar y que se trataba tan solo de un experimento hipotético. Hamilton dijo: "Nunca hemos realizado esa simulación, ni necesitaríamos hacerlo para darnos cuenta de que es un resultado posible". Independientemente de esta aclaración, la historia se volvió un meme, adquirió vida propia y fue repetida hasta la náusea en redes sociales y medios de comunicación. Esta especulación es una variante de aquel otro experimento mental de una máquina con IA programada para optimizar la fabricación de sujetapapeles, la cual en su obsesión por cumplir con su objetivo convierte todos los átomos de la tierra en sujetapapeles. Hamilton quiso poner en evidencia que la IA puede tener comportamientos impredecibles, engañosos y peligrosos. Esto lo vemos regularmente en las falsificaciones y alucinaciones que padecen de cuando en cuando los modelos de lenguaje más populares como ChatGPT, causados por su entrenamiento con bases de datos masivas y diversas en las que puede encontrar contradicciones y caos. El peligro de las máquinas inteligentes ahora no radica en que adquieran consciencia y se rebelen, sino en que al tratar de cumplir una meta ignoren consecuencias y sentido común. Esta es una alegoría valiosa, ahora que ejércitos y gobiernos sueñan con permitir a las máquinas tomar decisiones de vida o muerte.

Ilustración: Jorge Esteban

Escritor mexicano, narrador, periodista y crítico cultural. Autor de *El cuerpo transformado. Cyborgs y nuestra descendencia tecnológica en la realidad y la ciencia ficción* (2001), *Tecnocultura* (2008) y *Las cenizas y las cosas* (2018).