JAVIER CORTÉS, entrevista a Nick Bostrom: «No tendremos una segunda oportunidad con la inteligencia artificial», *El País*, 9 de desembre de 2017.

Nick Bostrom (Helsingborg, Suecia, 1973) es una de las voces más autorizadas para hablar de los peligros de los avances tecnológicos en nuestro tiempo. Este filósofo dirige el Instituto para el Futuro de la Humanidad y el Centro de Investigación de Estrategia de Inteligencia Artificial de la Universidad de Oxford, donde ejerce como profesor. Sus teorías sobre el riesgo que representaría la creación de una superinteligencia para el mundo han influido en el pensamiento de figuras como Bill Gates o Elon Musk.

Nos recibe en un pequeño camerino de la Fundación Giner de los Ríos, a unos pasos del escenario donde acaba de impartir una ponencia en el marco de un evento organizado por Adigital sobre cómo debemos afrontar la transición hacia la era de las máquinas inteligentes. Tiene que coger un vuelo en un par de horas y antes debe pasar por el hotel, pero se toma su tiempo antes de responder a cada pregunta y detiene su discurso varias veces para buscar las

palabras precisas. Es evidente que se encuentra cómodo especulando sobre el futuro que nos espera.

¿Cuáles son los principales riesgos tecnológicos con los que debe lidiar la humanidad?

Depende del contexto temporal que tengamos en mente. Si estamos pensando en las aplicaciones actuales, deberíamos estar preocupados por una serie de cosas que no tienen nada que ver con los peligros que vemos si miramos hacia el futuro, donde las máquinas tendrán niveles de inteligencia muy superiores a los de los humanos.

Hoy, uno de nuestros principales focos de atención apunta a que nuestros sistemas de información sean transparentes y a la vez nos permitan ciertos niveles de privacidad. Es muy importante que configuremos correctamente nuestra arquitectura informativa. Ahora nos interesa pensar en la forma en que se establecen las relaciones sociales en internet, la creación de filtros burbuja...

¿Y si echamos la vista al futuro?

Nos encontramos con un tipo de preocupaciones completamente distintas. No es difícil pensar en una inteligencia artificial que sea cada vez más poderosa cuyos objetivos no estén perfectamente alineados con los objetivos humanos. Y el reto aquí consiste en tener la tecnología que nos haga capaces de diseñar estos poderosos sistemas artificiales alineados con los valores humanos y que siempre hagan lo que nosotros queremos que hagan.

¿Es la inteligencia artificial nuestra única preocupación a largo plazo?

Existen otras áreas que también tienen un gran potencial riesgo para la humanidad. La biotecnología sería una de ellas. Si descubrimos cómo hacer mejores herramientas para manipular virus, bacterias y otros microrganismos, haríamos posible la creación de patógenos de diseño, que podría utilizar algún grupo terrorista para crear un nuevo tipo de enfermedad. Con cada nuevo avance técnico nos vamos aproximando a estos escenarios y, a medida que la tecnología madura, crece su potencial destructivo.

¿Considera que estamos avanzando demasiado rápido como sociedad?

Creo que cada vez que realizamos un gran descubrimiento, estamos metiendo la mano en una urna llena de pelotas y sacando una. Con cada pelota que sacamos, vamos alcanzando el progreso social. Toda la historia de la humanidad se ha basado en el proceso de hacer estos inventos.

La cuestión es que muchas de estas pelotas son beneficiosas, pero hay otras que pueden ser perjudiciales. Imagina lo que pasaría si una de las pelotas que sacamos es una tecnología que significa invariablemente la destrucción de la civilización que la descubrió. El problema aquí es que todavía no hemos descubierto una manera de volver a meter la pelota en la urna si no nos gusta: una vez que se ha inventado una tecnología, no puede «desinventarse».

¿Hemos sacado ya alguna de esas pelotas?

No es difícil adivinar cómo sería. Piensa, por ejemplo, en lo que pasaría si el desarrollo de armas nucleares hubiera sido posible con materiales más fáciles de conseguir. Si en

lugar de plutonio o uranio enriquecido, que no son precisamente accesibles, pudiera crearse una bomba atómica con un poco de arena, sería el fin de la civilización, porque bastarían unas pocas personas que quisieran causar daño para posibilitar una gran destrucción.

¿Podemos evitar que la inteligencia artificial termine siendo una de estas tecnologías?

Si la inteligencia artificial termina siendo capaz de hacer todo o buena parte de nuestro trabajo intelectual mejor que nosotros, tendremos en nuestras manos el último invento que tendrá que realizar la humanidad.

El problema está en la transición hasta la era de la inteligencia artificial: tenemos que hacerlo bien a la primera porque no creo que tengamos una segunda oportunidad. Si desarrollamos una inteligencia artificial que no esté alineada con nuestros propósitos, no creo que podamos volver a meter esa pelota en la urna y empezar de cero. No somos demasiado buenos como civilización anticipando problemas difíciles que todavía no nos han causado ningún daño. Debemos movilizar nuestros esfuerzos para hacer

que esto funcione desde el principio. Esta es la gran dificultad para nuestra civilización.

¿Qué tipo de políticas deben tomar los gobiernos y compañías tecnológicas al respecto?

Se nos presenta la oportunidad de acercarnos a la revolución de la inteligencia artificial de forma coordinada. Es el momento de trabajar en un proyecto internacional para desarrollar una superinteligencia segura en el que estén implicados todos los actores que puedan aportar algo. Si esto sale bien, los beneficios para la sociedad serían inmensos, pero podríamos sufrir consecuencias muy negativas si nuestra forma de abordar el problema consistiera en permitir que las empresas y naciones compitieran para ver quien consigue mejores resultados.

¿Cuál es la mejor manera de minimizar los riesgos de la IA?

Fundamentalmente, destinando recursos para desarrollar métodos de control que sean escalables, es decir, que sean más eficientes incluso si la IA se vuelve más inteligente. Por otra parte, deberíamos aprovechar el estado actual de esta tecnología para alinearla con valores

humanos. Hasta hace poco era un tema muy descuidado, pero cada vez hay más grupos de investigación trabajando en esta línea.

¿Estamos preparados para introducir valores humanos en una máquina?

Tenemos que definir los objetivos que queremos incorporar al sistema de IA para que interpreten estos valores como nos gustaría que fueran interpretados, y la manera de hacerlo es con el aprendizaje automático. No podemos especificar en un lenguaje informático lo que entendemos por justicia, placer o amor porque son conceptos muy complejos. En cambio, si programamos al sistema para que aprenda, podrá mejorar su comprensión de lo que quieren decir nuestras palabras en un sentido más profundo.