Uno de los ‘padres’ de la IA se incorpora a una iniciativa para prevenir sus peligros

Por: Melissa Heikkilä. 23/08/2024

El ganador del Premio Turing cree que la única forma de garantizar la seguridad de los sistemas de IA es utilizar la propia IA en lugar de humanos

Yoshua Bengio, ganador del premio Turing y considerado **uno de los “padrinos” de la IA moderna**, apoya un proyecto financiado por el gobierno británico para incorporar **mecanismos de seguridad** a los sistemas de IA.

El proyecto, denominado Safeguarded AI (IA salvaguardada, en inglés), pretende crear un **sistema de IA capaz de comprobar si otros sistemas de IA desplegados en zonas críticas son seguros.** Bengio se incorpora al programa como director científico y aportará información crítica y asesoramiento científico. El proyecto, que recibirá 59 millones de libras en los próximos cuatro años, está financiado por la Agencia de Investigación e Invención Avanzadas del Reino Unido (ARIA, por sus siglas en inglés), creada en enero del año pasado para invertir en investigación científica potencialmente transformadora.

El objetivo de Safeguarded AI es construir **sistemas de IA que puedan ofrecer garantías cuantitativas**, como una puntuación de riesgo, sobre su efecto en el mundo real, explica David “davidad” Dalrymple, director del programa Safeguarded AI en ARIA. La idea es complementar las pruebas humanas con análisis matemáticos del potencial de daño de los nuevos sistemas.

El proyecto pretende crear mecanismos de seguridad de la IA combinando modelos científicos del mundo, que son esencialmente simulaciones del mundo, con pruebas matemáticas. Estas pruebas incluirían explicaciones del trabajo de la IA, y los humanos se encargarían de verificar si las comprobaciones de seguridad del modelo de IA son correctas.

Bengio afirma que quiere contribuir a garantizar **que los futuros sistemas de IA no puedan causar daños graves.**

“Actualmente estamos corriendo hacia una niebla tras la cual podría haber un precipicio”, afirma. “No sabemos a qué distancia está el precipicio, o si siquiera existe, así que podrían pasar años, décadas, y no sabemos lo grave que podría ser… Tenemos que crear las herramientas para despejar esa niebla y asegurarnos de que no cruzamos el precipicio, si es que existe”.

Las empresas científicas y tecnológicas no tienen forma de dar garantías matemáticas de que los sistemas de IA vayan a comportarse según lo programado, añade. Esta falta de fiabilidad, según afirma, podría provocar **resultados catastróficos.**

Dalrymple y Bengio sostienen que las técnicas actuales para mitigar el riesgo de los sistemas avanzados de IA, como el “red-teaming” [equipo independiente encargado de atacar las vulnerabilidades], en el que unas personas examinan los sistemas de IA en busca de fallos, tienen serias limitaciones y no se puede confiar en ellas para garantizar que los sistemas críticos no se salgan de pista.

En su lugar, **esperan que el programa proporcione nuevas formas de hacer seguros los sistemas de IA** que dependan menos de los esfuerzos humanos y más de la certeza matemática. La idea es **construir una IA “guardiana”** que se encargue de comprender y reducir los riesgos de seguridad de otros agentes de IA. Este guardián garantizaría que los agentes de IA que trabajan en sectores de alto riesgo, como el transporte o los sistemas energéticos, funcionen como queremos. La idea es colaborar con las empresas desde el principio para entender cómo los mecanismos de seguridad de la IA podrían ser útiles para distintos sectores, afirma Dalrymple.

La complejidad de los sistemas avanzados hace que no tengamos más remedio que utilizar la IA para salvaguardarla, argumenta Bengio. “Es la única manera, porque llega un momento en que **estas IA son demasiado complicadas**. Incluso con las que ya tenemos no podemos descomponer sus respuestas en secuencias de razonamiento humanas y comprensibles”, afirma.

El siguiente paso, la construcción de modelos que puedan comprobar otros sistemas de IA, es también donde Safeguarded AI y ARIA esperan cambiar el *statu quo* de la industria de la IA.

ARIA también ofrece financiación a personas u organizaciones de sectores de alto riesgo como el transporte, las telecomunicaciones, las **cadenas de suministro** y la investigación médica para ayudarles a crear aplicaciones que puedan beneficiarse de los mecanismos de seguridad de la IA. ARIA ofrece a los solicitantes un total de 5,4 millones de libras el primer año, y otros 8,2 millones el siguiente. El plazo de presentación de solicitudes finaliza el 2 de octubre.

La agencia también está lanzando una amplia red para personas que puedan estar interesadas en construir el mecanismo de seguridad de la IA salvaguardada a través de una organización sin ánimo de lucro. ARIA aspira a conseguir **hasta 18 millones de libras** para crear esta organización y aceptará solicitudes de financiación a principios del año que viene.

El programa busca propuestas para poner en marcha una organización sin ánimo de lucro con un consejo diverso que abarque muchos sectores diferentes para poder realizar este trabajo de forma fiable y fidedigna, afirma Dalrymple. Esto es similar a lo que OpenAI se propuso hacer inicialmente antes de cambiar su estrategia y orientarse más hacia los productos y los beneficios.

El consejo de la organización no sólo se encargará de pedir cuentas al director general, sino que incluso influirá en las decisiones sobre si **emprender o no determinados proyectos de investigación** y si publicar o no determinados documentos y API, añade.

El proyecto Safeguarded AI forma parte de la misión del Reino Unido de **posicionarse como pionero en seguridad de la IA.** En noviembre de 2023, el país acogió la primera Cumbre sobre Seguridad de la IA, que reunió a líderes mundiales y tecnólogos para debatir cómo construir la tecnología de forma segura.

Aunque el programa de financiación tiene preferencia por los solicitantes con sede en el Reino Unido, ARIA busca talentos globales que puedan estar interesados en venir al Reino Unido, dice Dalrymple. ARIA también cuenta con un mecanismo de propiedad intelectual para financiar empresas con ánimo de lucro en el extranjero, que permite que los derechos de autor retornen al país.

Bengio afirma que se sintió atraído por el proyecto para fomentar la colaboración internacional en materia de seguridad de la IA. Preside el Informe Científico

Internacional sobre la seguridad de la IA avanzada, en el que participan 30 países, además de la UE y la ONU. Defensor a ultranza de la seguridad de la IA, ha formado parte de un influyente grupo de presión que advierte de que la **IA superinteligente** supone un **riesgo existencial.**

“Tenemos que llevar el debate sobre cómo vamos a abordar los riesgos de la IA a un conjunto global y más amplio de actores”, afirma Bengio. “Este programa nos acerca a ello”.

LEER EL ARTÍCULO ORIGINAL PULSANDO AQUÍ

Fotografía: Technology review

Fecha de creación
2024/08/23