El 'Padrino de la IA' ahora teme que no sea seguro. Tienen un plan para controlarlo

theconversation.com/godfather-of-ai-now-fears-its-unsafe-he-has-a-plan-to-rein-it-in-258288

Armin Chitizadeh

Esta semana, la Oficina Federal de Investigaciones de Estados Unidos (FBI, por sus siglas en inglés) reveló que dos hombres sospechosos de bombardear una clínica de fertilidad en California el mes pasado presuntamente utilizaron inteligencia artificial (IA) para obtener instrucciones para fabricar bombas. El FBI no reveló el nombre del programa de IA en cuestión.

Esto pone de manifiesto la urgente necesidad de hacer que la IA sea más segura. Actualmente vivimos en la era del "salvaje oeste" de la IA, donde las empresas compiten ferozmente para desarrollar los sistemas de IA más rápidos y entretenidos. Cada empresa quiere superar a sus competidores y reclamar el primer puesto. Esta intensa competencia a menudo conduce a atajos intencionales o no intencionales , especialmente cuando se trata de seguridad.

Casualmente, casi al mismo tiempo que la revelación del FBI, uno de los padrinos de la IA moderna, el profesor canadiense de ciencias de la computación Yoshua Bengio, lanzó una nueva organización sin fines de lucro dedicada a desarrollar un nuevo modelo de IA diseñado específicamente para ser más seguro que otros modelos de IA y apuntar a aquellos que causan daño social.

Entonces, ¿cuál es el nuevo modelo de IA de Bengio? ¿Y realmente protegerá al mundo de los daños facilitados por la IA?

Una IA "honesta"

En 2018, Bengio, junto con sus colegas Yann LeCun y Geoffrey Hinton, ganó el Premio Turing por una investigación pionera que habían publicado tres años antes sobre el aprendizaje profundo. El aprendizaje profundo, una rama del aprendizaje automático, intenta imitar los procesos del cerebro humano mediante el uso de redes neuronales artificiales para aprender de los datos computacionales y hacer predicciones.

La nueva organización sin ánimo de lucro de Bengio, LawZero, está desarrollando "Scientist AI". Bengio ha dicho que este modelo será "honesto y no engañoso", e incorporará principios de seguridad por diseño .

Según un documento preimpreso publicado en línea a principios de este año, Scientist AI diferirá de los sistemas de IA actuales en dos aspectos clave.

En primer lugar, puede evaluar y comunicar su nivel de confianza en sus respuestas, lo que ayuda a reducir el problema de que la IA dé respuestas demasiado seguras e incorrectas.

En segundo lugar, puede explicar su razonamiento a los humanos, lo que permite que sus conclusiones sean evaluadas y probadas para determinar su precisión.

Curiosamente, los tallos AIs y más antiguos tenían esta característica. Pero en la prisa por la velocidad y los nuevos enfoques, muchos modelos modernos de IA no pueden explicar sus decisiones. Sus desarrolladores han sacrificado la explicabilidad por la velocidad.

Bengio también tiene la intención de que "Scientist AI" actúe como una barandilla contra la IA insegura. Podría monitorear otros sistemas de IA menos confiables y dañinos, esencialmente combatiendo el fuego con fuego.

Esta puede ser la única solución viable para mejorar la seguridad de la IA. Los humanos no pueden monitorear adecuadamente sistemas como ChatGPT, que manejan más de mil millones de consultas diarias. Solo otra IA puede manejar esta escala.

El uso de un sistema de IA frente a otros sistemas de IA no es solo un concepto de ciencia ficción, sino que es una práctica común en la investigación comparar y probar diferentes niveles de inteligencia en los sistemas de IA.

Añadiendo un "modelo mundial"

Los grandes modelos de lenguaje y el aprendizaje automático son solo pequeñas partes del panorama actual de la IA.

Otra adición clave que el equipo de Bengio está agregando a Scientist AI es el "Modelo mundial" que aporta certeza y explicabilidad. Al igual que los humanos toman decisiones basadas en su comprensión del mundo, la IA necesita un modelo similar para funcionar de manera efectiva. La ausencia de un modelo mundial en los modelos actuales de IA es evidente.

Un ejemplo bien conocido es el "Problema con la mano": la mayoría de los modelos de IA actuales pueden imitar la apariencia de las manos, pero no pueden replicar los movimientos naturales de las manos, porque carecen de una comprensión de la física, un modelo mundial, detrás de ellos.

Otro ejemplo es cómo modelos como ChatGPT se estruendan con el ajedrez, fracasan para ganar e incluso hacen jugadas.

Esto a pesar de que los sistemas de IA más simples, que contienen un modelo del "mundo" del ajedrez, superan incluso a los mejores jugadores humanos.

Estos problemas se derivan de la falta de un modelo mundial fundamental en estos sistemas, que no son inherentes a modelar la dinámica del mundo real.

Yoshua Bengio es reconocido como uno de los padrinos de la IA. Alex Wong/Getty Images

En el camino correcto, pero estará lleno de baches

Bengio va por buen camino, con el objetivo de crear una IA más segura y fiable mediante la combinación de grandes modelos lingüísticos con otras tecnologías de IA.

Sin embargo, su camino no va a ser fácil. Los 30 millones de dólares de LawZero son pequeños en comparación con esfuerzos como el proyecto de 500.000 millones de dólares anunciado por el presidente de Estados Unidos, Donald Trump, a principios de este año para acelerar el desarrollo de la IA.

Lo que dificulta la tarea de LawZero es el hecho de que Scientist AI, como cualquier otro proyecto de IA, necesita enormes cantidades de datos para ser potente, y la mayoría de los datos están controlados por empresas tecnológicas o por empresas.

También hay una pregunta pendiente. Incluso si Bengio puede construir un sistema de IA que haga todo lo que dice que puede, ¿cómo va a ser capaz de controlar otros sistemas que podrían estar causando daño?

Aun así, este proyecto, con investigadores talentosos detrás, podría desencadenar un movimiento hacia un futuro en el que la IA realmente ayude a los humanos a prosperar. Si tiene éxito, podría establecer nuevas expectativas para una IA segura, motivando a investigadores, desarrolladores y responsables políticos a priorizar la seguridad.

Tal vez si hubiéramos tomado medidas similares cuando surgieron las redes sociales, tendríamos un entorno en línea más seguro para la salud mental de los jóvenes. Y tal vez, si Scientist AI ya hubiera estado en su lugar, podría haber evitado que personas con intenciones dañinas accedieran a información peligrosa con la ayuda de sistemas de IA.