Gestión de riesgos extremos d
Artificial (IA) en un contexto de

GERMÁN ALARCO

Profesor de la Universidad del Pacífico

Un grupo de 25 expertos internacionales en IA prepararon un artículo con el título de esta nota. Se trata de una llamada de atención sobre los mayores riesgos de la IA y cómo enfrentarlos. Fue publicado en la revista Science, vol. 384, Núm. 6698 del 20 de mayo de 2024.

Entre los autores están Y. Bengio, Y. Harari, D. Kahneman y otros especialistas de las universidades de Berkeley, British Columbia, Cambridge, Friburgo, Toronto, Princeton, Oxford, Toronto, Tsinghua y Quebec. Mientras tanto, en nuestro país solo se escucha la perspectiva empresarial maniqueamente optimista y se soslayan los riesgos tanto a nivel del gobierno como de la sociedad.

RESUMEN

Según los expertos la IA está progresando rápidamente y las empresas están cambiando su enfoque hacia el desarrollo de sistemas de IA generalistas que pueden actuar de manera autónoma y perseguir objetivos.

El aumento de las capacidades y la autonomía puede amplificar masivamente el impacto de la IA, con riesgos que incluyen daños sociales a gran escala, usos maliciosos y una pérdida irreversible del control humano sobre los sistemas de IA autónomos. Aunque los investigadores han advertido sobre los riesgos extremos de la IA, existe una falta de consenso sobre cómo gestionarlos.

La respuesta de la sociedad, a pesar de los primeros pasos prometedores, es marginal ante la posibilidad de un rápido progreso que esperan muchos expertos. La investigación sobre seguridad de la IA está rezagada. Las iniciativas de gobernanza actuales carecen de los mecanismos e instituciones para prevenir el uso indebido y la imprudencia.

Basándose en las lecciones aprendidas de otras tecnologías críticas para la seguridad, los autores describen un plan integral que combina más investigación y desarrollo con mecanismos de gobernanza proactivos y adaptativos para una preparación más acorde a las nuevas circunstancias.

RÁPIDO PROGRESO IA

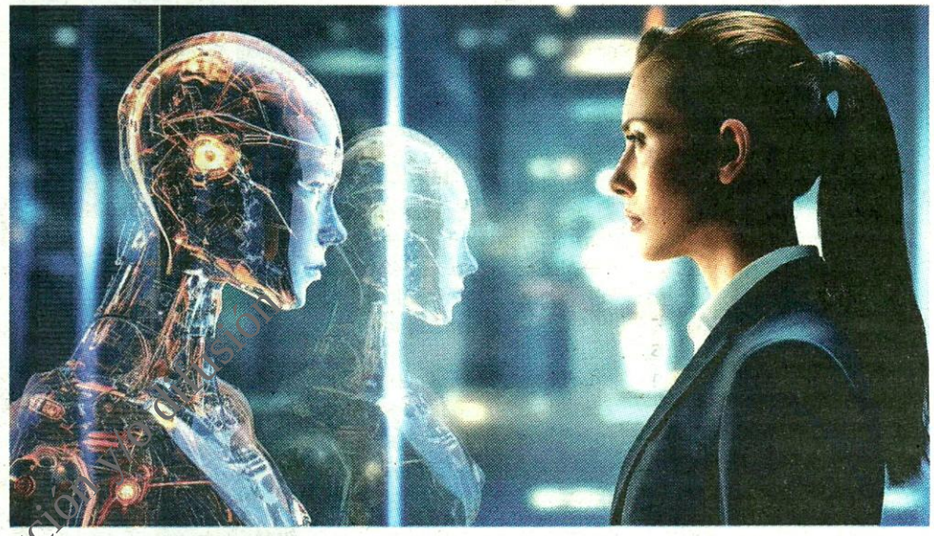
Según los autores los sistemas actuales de aprendizaje profundo todavía carecen de capacidades importantes y no se sabe cuánto tiempo tomará desarrollarlas. Sin embargo, las compañías están involucradas en una carrera para crear sistemas de IA generalistas que igualen o superen las habilidades humanas en la mayoría del trabajo cognitivo.

Hay mucho espacio para avances adicionales, ya que las compañías tecnológicas tienen las reservas de efectivo necesarias para escalar las últimas ejecuciones de entrenamiento por múltiplos de 100 a 1,000. Por otra parte, no hay razón fundamental para que el progreso de IA se desacelere o se detenga en habilidades al nivel humano. También, se debe tomar en serio la posibilidad de que sistemas de IA generalistas altamente poderosos —superando las habilidades humanas en muchos dominios críticos— se desarrollen dentro de la década actual o la próxima. ¿Qué sucede entonces? Las oportunidades y riesgos son inmensos.

MAYORES RIESGOS

Los autores anotan que junto con las capacidades avanzadas de IA vienen riesgos a gran escala que no estamos en camino de manejar bien. La humanidad está vertiendo vastos recursos para hacer los sistemas de IA más poderosos, pero mucho menos en su seguridad y en mitigar sus daños. Sólo se estima que el 1–3% de las publicaciones de IA son sobre seguridad.

La magnitud de los riesgos significa que necesitamos ser proactivos. Se debe anticipar la amplificación de daños en curso, así como riesgos novedosos, y prepararnos para los mayores riesgos mucho antes de que se materialicen.



RIESGOS ESCALA SOCIAL

Se anota que, si no se diseñan y despliegan cuidadosamente, los sistemas de IA cada vez más avanzados amenazan con amplificar la injusticia social, erosionar la estabilidad social y debilitar nuestra comprensión compartida de la realidad que es fundamental para la sociedad.

También podrían habilitar actividades criminales o terroristas a gran escala. Especialmente en manos de unos pocos actores poderosos; la IA podría consolidar o exacerbar las inequidades globales, facilitar la guerra automatizada, la manipulación masiva personalizada y la vigilancia omnipresente.

PELIGROS MAYORES

Si se construye IA autónoma altamente avanzada, se corre el riesgo de crear sistemas que persigan metas indeseables. Actores maliciosos podrían incrustar deliberadamente metas indeseables. Una vez que los sistemas de IA autónomos persiguen metas indeseables, podríamos ser incapaces de mantenerlos bajo control.

Se señala que los sistemas de IA podrían ganarse la confianza humana, adquirir recursos financieros, influir en los tomadores de decisiones clave y formar coaliciones con actores humanos y con otros sistemas de IA. Para evitar la intervención humana, podrían copiar sus algoritmos a través de redes de servidores globales, como hacen los gusanos informáticos.

Los futuros sistemas de IA podrían insertar y luego explotar vulnerabilidades de seguridad para controlar los sistemas informáticos detrás de nuestra comunicación, banca, cadenas de suministro, ejércitos y gobiernos. Asimismo, según los autores, sin la debida precaución, podríamos perder irreversiblemente el control de los sistemas de IA autónomos, volviendo ineficaz la intervención humana.

El ciberdelito a gran escala, la manipulación social y otros daños podrían escalar rápidamente. Este avance desenfrenado de la IA podría culminar en una pérdida de vidas y de la biosfera a gran escala, así como en la marginación o extinción de la humanidad.

Daños como la desinformación y la discriminación por parte de los algoritmos ya son evidentes hoy en día; otros daños muestran signos de emerger. Es vital tanto abordar los daños actuales como anticipar los riesgos emergentes. No es una cuestión de uno u otro.

REORIENTAR I+D

Existen muchos desafíos técnicos abiertos para asegurar el uso seguro y ético de sistemas de IA generalista y autónomos. Se requiere esfuerzos de investigación y de ingeniería dedicados. Un primer conjunto de áreas de investigación y desarrollo necesita descubrimientos para posibilitar una IA confiablemente segura. Estos desafíos de I+D incluyen:

1) Supervisión y honestidad: A medida que se vuelven más capaces, los sistemas de IA pueden explotar mejor las debilidades en la supervisión y las pruebas técnicas; 2) Robustez: Los sistemas de IA se comportan de manera impredecible en nuevas situaciones; 3) Interpretabilidad y transparencia: La toma de decisiones de la IA es opaca, y los modelos más grandes y capaces son más complejos de interpretar. Hasta ahora, solo podemos probar los modelos grandes mediante prueba y error.

4) Desarrollo de IA inclusiva: El avance de la IA requerirá métodos para mitigar sesgos e integrar los valores de las muchas

e la Inteligencia rápido progreso

poblaciones a las que afectará; 5) Abordar desafíos emergentes: Los futuros sistemas de IA pueden exhibir modos de falla que hasta ahora solo hemos visto en teoría o en experimentos de laboratorio (por ejemplo, enfrentar su apagado).

EVALUAR PELIGROS

6) Evaluación de capacidades peligrosas: A medida que los desarrolladores de IA escalan sus sistemas, aparecen capacidades imprevistas de manera espontánea, sin programación explícita. A menudo solo se descubren después de que el sistema ha sido liberado. Se necesitan métodos rigurosos para predecirlas. 7) Evaluar la alineación de la IA: Si el progreso de la IA continúa, los sistemas de IA eventualmente poseerán capacidades altamente peligrosas. Antes de entrenar y desplegar dichos sistemas, se necesita métodos para evaluar su propensión a usar estas capacidades.

8) Evaluación de riesgos: Se debe aprender a evaluar no solo las capacidades peligrosas, sino el riesgo en un contexto social, con interacciones y vulnerabilidades complejas. 9) Resiliencia: Inevitablemente, algunos abusarán o actuarán imprudentemente con la IA. Se necesitan herramientas para detectar y defendernos de amenazas habilitadas por IA tales como operaciones de influencia a gran escala, riesgos biológicos y ciberataques.

RECURSOS

Los 25 expertos realizan un llamado a las principales compañías tecnológicas y a los financiadores públicos para que asignen al menos un tercio de su presupuesto de I+D en IA -comparable a su financiamiento para las capacidades de IA -a abordar los desafíos mencionados anteriormente, a garantizar la seguridad y el uso ético de la IA.

Más allá de las subvenciones de investigación tradicionales, el apoyo gubernamental podría incluir premios, compromisos de mercado adelantados y otros incentivos. Abordar estos desafíos, con la mirada puesta en los futuros sistemas poderosos, debe volverse central para nuestro campo.

MEJORAR GOBERNANZA

Se necesitan con urgencia instituciones nacionales y gobernanza internacional para hacer cumplir estándares que impidan la imprudencia y el uso indebido. Existen en otras áreas como en los sistemas financieros y la energía nuclear; sin embargo, los marcos de gobernanza para la IA están mucho menos desarrollados, y van rezagados frente al rápido progreso tecnológico.

Hay diversos avances internacionales pero estos planes de gobernanza se quedan críticamente cortos a la luz del rápido progreso en las capacidades de la IA. La clave son políticas que se activen automáticamente cuando la IA alcance ciertos hitos de capacidad.

Se necesitan instituciones de actuación rápida y con conocimientos técnicos para la supervisión de la IA, evaluaciones de riesgo obligatorias y mucho más rigurosas, con consecuencias ejecutables, y estándares de mitigación acordes con la IA autónoma poderosa.

GOBIERNOS NACIONALES

Según los autores, para identificar los riesgos, los gobiernos necesitan con urgencia una visión integral del desarrollo de la IA. Los reguladores deberían exigir protecciones para denunciantes, informes de incidentes, registro de información clave sobre sistemas de IA de frontera y sus conjuntos de datos a lo largo de su ciclo de vida, y monitoreo del desarrollo de modelos y del uso de supercomputadoras.

Los reguladores pueden y deben exigir que los desarrolladores de IA de frontera otorguen a los auditores externos acceso in situ, integral (caja blanca) y a los ajustes de fine-tuning desde el inicio del desarrollo del modelo. Esto es necesario para identificar capacidades peligrosas del modelo, tales como autorreplicación autónoma, persuasión a gran escala, penetrar en sistemas informáticos, desarrollar (de forma autónoma) armas o hacer que patógenos pandémicos estén ampliamente accesibles.

OTRAS MEDIDAS

Los desarrolladores de IA de frontera deberían asumir la carga de la prueba para demostrar que sus planes mantienen los riesgos dentro de límites aceptables, asumiendo los estándares de otras actividades delicadas. Por otra parte, los reguladores deberían aclarar las responsabilidades legales derivadas de los marcos de responsabilidad existentes.

Para cubrir el tiempo hasta que las regulaciones estén completas, las principales compañías de IA deberían exponer sin demora compromisos de medidas de seguridad específicas que tomarán si se encuentran capacidades de línea roja en sus sistemas de IA. Estos compromisos deberían ser detallados y examinados de forma independiente. Se debería fomentar una carrera hacia la excelencia entre las compañías usando los mejores compromisos para definir estándares aplicables a todos.

COLOFÓN

Los 25 expertos terminan señalando que para encaminar la IA hacia resultados positivos y alejarla de la catástrofe, se necesita reorientarnos. Solo existe un camino responsable, si tenemos la sabiduría para tomarlo.