Universidad del Pacífico Universidad del Pacífico	Medio Impreso		Página	Fecha
	Diario Uno		8	27/12/2024
			Valorización	Medida (cm2)
			US $ 4,480.30	773.21 (86.00%)

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.

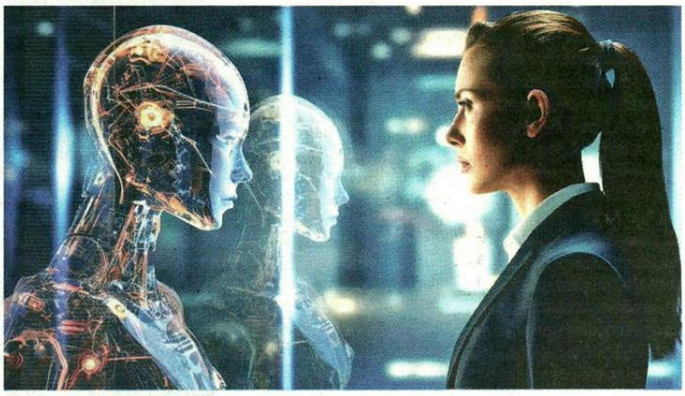
La magnitud de los riesgos significa que necesitamos ser proactivos. Se debe anticipar la amplificación de daños en curso, así como riesgos novedosos, y prepararnos para los mayores riesgos mucho antes de que se materialicen.

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