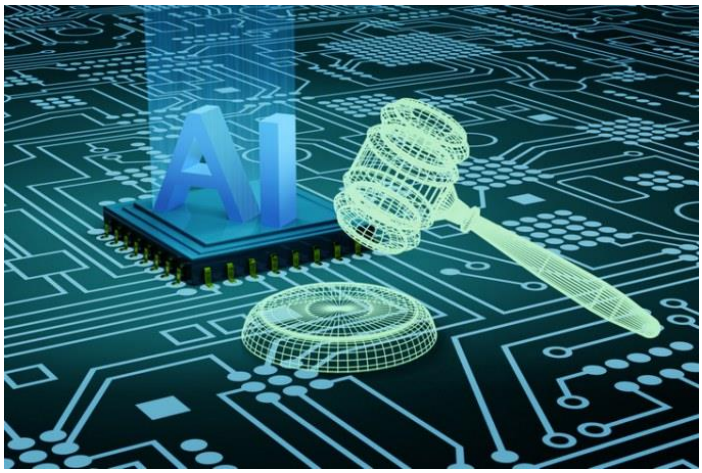
Link para original:
https://www.prevencionintegral.com/actualidad/noticias/2024/05/23/grupo-expertos-presenta-documento-para-gestion-riesgos-extremos-que-plantea-ia?utm_source=cerpie&utm_medium=email&utm_campaign=flash_300524

# Un grupo de expertos presenta un documento para la gestión de los riesgos extremos que plantea la IA

*La IA está experimentando una rápido avance que conduce al desarrollo que sistemas cada vez más capaces y autónomos que plantean cada vez mayores riesgos sociales. Abordarlos, requiere tanto medidas de desarrollo técnico como de gobernanza.*

30 Mayo 2024




La revista 'Science' ha publicado un documento en el que quince expertos de primer nivel mundial en materia de Inteligencia Artificial (IA), presentan unas directrices ante los riesgos extremos que plantea el rápido avance de este tecnología.

El documento "Managing extreme AI risks amid rapid progress" (https://www.science.org/doi/10.1126/science.adn0117) está impulsado por Yoshua Bengio y firmado por figuras tan relevantes del sector como Geoffrey Hinton, Trevor Darrell, Yuval Noah Harari, Jeff Clune, Sheila McIlraith, David Krueger, Philip Torr, Stuart Russell, Sören Mindermann y el recientemente fallecido Daniel Kahneman, a cuya memoria se dedica. Muchas de estos firmantes ya lo fueron de otro documento previo, difundido tras la aparición de ChatGPT, en el que se solicitaba una moratoria en el desarrollo de estos sistemas, por razones muy similares.

En este nuevo escrito se advierte de que la IA está progresando rápidamente, impulsada por una actividad de las principales empresas del sector que se está orientado hacia el desarrollo de sistemas de IA generalistas que puedan actuar y perseguir objetivos de forma autónoma.

De esta forma, los firmantes consideran que el incremento de las capacidades y la autonomía de estos sistemas pronto podrían amplificar enormemente el impacto de la IA, planteando unos riesgos que incluyen daños sociales a gran escala, usos maliciosos y una pérdida irreversible del control humano sobre los sistemas autónomos de IA.

Y aunque numerosos investigadores han advertido ya sobre los riesgos extremos que plantean estos desarrollos, no existe consenso sobre cómo gestionarlos.

En este contexto, los expertos firmantes del documento sostienen que a pesar de unos primeros pasos prometedores, la respuesta de la sociedad a estos desafíos, no es proporcionada a la rapidez con que avanza esta tecnología. La investigación sobre seguridad de la IA está retrasada y las iniciativas de gobernanza actuales carecen de mecanismos e instituciones para prevenir el mal uso de esta tecnología y apenas abordan los sistemas autónomos.

Por ello, y basándose en las lecciones aprendidas de otras tecnologías críticas para la seguridad, estos expertos proponen un plan integral que combina **investigación y desarrollo técnicos (I+D)** con mecanismos de gobernanza proactivos y adaptables para una preparación más acorde para afrontar estos riesgos.

En concreto, se plantea el establecimiento de una serie de medidas proporcionadas para futuros sistemas de IA excepcionalmente capaces, como sistemas autónomos que podrían eludir el control humano.

Según estos expertos, los gobiernos deben estar preparados para autorizar su desarrollo, restringir su autonomía en roles sociales clave, detener su desarrollo y despliegue en respuesta a capacidades preocupantes, exigir controles de acceso y establecer medidas de seguridad de la información sólidas contra los piratas informáticos a nivel estatal hasta que estén listas las protecciones adecuadas. Los gobiernos deberían desarrollar estas capacidades ahora.

Para acortar el tiempo hasta que se adopte esa regulación, las principales empresas de IA deberían establecer rápidamente compromisos if-then ("si-entonces"), como medidas de seguridad específicas a tomar en caso de que los equipos de seguridad (Red Teams), identifiquen determinadas capacidades específicas. Estos compromisos deben ser detallados y examinados de forma independiente.

Estas medidas se centran en:

1. REORIENTAR LA I+D TÉCNICA

Según los autores del documento, garantizar la seguridad y el uso ético de los sistemas de IA autónomos y generalistas plantea muchos desafíos técnicos. Pero, a diferencia de lo que ocurre con el avance de las capacidades de la IA, estos desafíos no pueden abordarse simplemente utilizando más potencia informática para entrenar modelos más grandes. Por tanto, es poco probable que se resuelvan automáticamente a medida que los sistemas de IA se vuelvan más capaces y requieren esfuerzos dedicados de investigación e ingeniería. Por lo tanto, no sabemos si el trabajo técnico puede resolver fundamentalmente estos desafíos a tiempo. Sin embargo, se ha trabajado comparativamente poco en muchos de estos desafíos. Por tanto, una mayor I+D puede facilitar el progreso y reducir los riesgos.

Un primer conjunto de áreas de I+D necesita avances que permitan una IA segura y fiable. Sin estos avances, los desarrolladores deben arriesgarse a crear sistemas inseguros o a quedarse rezagados frente a competidores dispuestos a asumir más riesgos. Si garantizar la seguridad sigue siendo demasiado difícil, se necesitarían medidas de gobernanza extremas para evitar los recortes impulsados por la competencia y el exceso de confianza. Entre estos retos de I+D figuran los siguientes:

- Supervisión y honestidad: Los sistemas de IA más capaces pueden explotar mejor las debilidades en la supervisión y las pruebas técnicas, por ejemplo, produciendo resultados falsos pero convincentes.

- Robustez: Los sistemas de IA se comportan de forma impredecible en situaciones nuevas. Mientras que algunos aspectos de la robustez mejoran con la escala del modelo, otros aspectos no mejoran o incluso empeoran.

- 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, los modelos más grandes sólo se pueden probar mediante el sistema de prueba y error. Por ello es necesario aprender a comprender su funcionamiento interno.

- Desarrollo inclusivo de la IA: El avance de la IA necesitará métodos para mitigar los sesgos e integrar los valores de las muchas poblaciones a las que afectará.

- Abordar los desafíos emergentes: Los futuros sistemas de IA pueden presentar fallos que hasta ahora solo hemos visto en teoría o experimentos de laboratorio, como que los sistemas de IA tomen el control de los canales de provisión de recompensas de capacitación o exploten las debilidades en nuestros objetivos de seguridad y mecanismos de parada para avanzar en un objetivo particular.

Un segundo conjunto de desafíos de I+D necesita de avances para permitir una gobernanza eficaz y ajustada al riesgo o para reducir los daños cuando la seguridad y la gobernanza fallan. En este sentido, se destaca la necesidad de:

- Evaluar las capacidades peligrosas de la IA: A medida que los desarrolladores de IA escalan sus sistemas, aparecen capacidades imprevistas de forma espontánea, sin programación explícita. A menudo, estos casos sólo se descubren después de la implementación. Por ello se necesitan métodos rigurosos para obtener y evaluar las capacidades de la IA y predecirlas antes del entrenamiento. Esto incluye tanto capacidades genéricas para lograr objetivos ambiciosos (por ejemplo, planificación y ejecución a largo plazo), como capacidades peligrosas específicas basadas en modelos de amenazas (por ejemplo, manipulación social o piratería informática). Las evaluaciones actuales sobre las capacidades peligrosas de los modelos de IA más avanzadas (o de frontera), que son clave para diversos marcos de políticas de IA, se limitan a controles aleatorios e intentos de prueba en entornos específicos. Estas evaluaciones a veces pueden demostrar capacidades peligrosas de los sistemas de IA, pero no permiten descartarlas de manera confiable: los sistemas de IA que carecieron de ciertas capacidades en las pruebas pueden presentarlas en entornos ligeramente diferentes o con mejoras posteriores al entrenamiento. Por lo tanto, las decisiones que dependen de que los sistemas de IA no crucen ninguna línea roja necesitan grandes márgenes de seguridad. Las herramientas de evaluación mejoradas reducen la

posibilidad de pasar por alto capacidades peligrosas, lo que permite márgenes más pequeños.

- Evaluar la alineación de la IA: Si el progreso de la IA continúa, los sistemas de IA podrían alcanzar eventualmente capacidades altamente peligrosas. Por tanto, antes de entrenar y desplegar dichos sistemas, son necesarios métodos para evaluar su propensión a utilizar estas capacidades. Las evaluaciones puramente conductuales pueden fallar para los sistemas avanzados de IA: al igual que los humanos, pueden comportarse de manera diferente bajo evaluación, fingiendo alineación.

- Evaluación de riesgos: Debemos aprender a evaluar no sólo las capacidades peligrosas sino también el riesgo en un contexto social, con interacciones y vulnerabilidades complejas. La evaluación rigurosa de riesgos para los sistemas de IA de vanguardia sigue siendo un desafío abierto debido a sus amplias capacidades y su despliegue generalizado en diversas áreas de aplicaciones.

- Resiliencia: Inevitablemente, algunos harán mal uso de la IA o actuarán imprudentemente. Por ello se necesitan herramientas para detectar y defendernos contra amenazas habilitadas por la IA, como operaciones de influencia a gran escala, riesgos biológicos y ataques cibernéticos. Y como resulta, además que, a medida que los sistemas de IA se vuelvan más capaces, podrían eventualmente eludir las defensas creadas por los humanos, primero debemos aprender cómo hacer que los sistemas de IA sean seguros y estén alineados.

En resumen, teniendo en cuenta lo que está en juego, los firmantes del documento hacen un llamado a las principales empresas de tecnología y a los financiadores públicos para que asignen al menos un tercio de su presupuesto de I+D en IA, comparable a su financiación para capacidades de IA, para abordar los desafíos de I+D mencionados anteriormente y garantizar la seguridad y el uso ético de la IA. Más allá de las tradicionales subvenciones para investigación, el apoyo gubernamental podría incluir premios, compromisos anticipados de mercado y otros incentivos. Abordar estos desafíos, con miras a sistemas futuros poderosos, debe convertirse en algo central en nuestro campo.

2. MEDIDAS DE GOBERNANZA

Los firmantes del documento subrayan en este sentido que se necesita urgentemente instituciones nacionales y de gobernanza internacional para hacer cumplir normas que prevengan la imprudencia y el mal uso de los sistemas de IA. Muchas áreas de la tecnología, desde los productos farmacéuticos hasta los sistemas financieros y la energía nuclear, muestran que la sociedad requiere y utiliza efectivamente la supervisión gubernamental para reducir los riesgos. Sin embargo, los marcos de gobernanza para la IA están mucho menos desarrollados y van a la zaga del rápido progreso tecnológico. Por ello, proponen inspirarnos en la gobernanza de otras tecnologías críticas para la seguridad y al mismo tiempo tener en cuenta las características distintivas de la IA avanzada, como son su capacidad de superar con creces a otras tecnologías en su potencial para actuar y desarrollar ideas de forma autónoma, progresar explosivamente, comportarse de manera adversaria y causar un daño irreversible.

Los firmantes reconocen que gobiernos de todo el mundo han dado pasos positivos en relación con los sistemas de IA más avanzados. Actores clave como China, Estados

Unidos, la Unión Europea y el Reino Unido han elaborado directrices o regulaciones iniciales al respecto que, pese a sus limitaciones (carácter voluntario, alcance geográfico limitado y exclusión de áreas de alto riesgo como sistemas militares y en fase de I+D), son pasos iniciales importantes para establecer, entre otros aspectos, temas como la responsabilidad de los desarrolladores, las auditorías de terceros y los estándares de la industria.

Sin embargo, estos planes de gobernanza resultan críticamente insuficientes en vista del rápido progreso en las capacidades de IA. En este contexto son necesarias medidas de gobernanza que nos preparen para avances repentinos en la IA y que al mismo tiempo sean políticamente viables, a pesar de los desacuerdos y la incertidumbre sobre los cronogramas de la IA. La clave son las políticas que se activen automáticamente cuando la IA alcance determinados hitos de capacidad, de forma que si la IA alcanza un nivel de desarrollo suficientemente rápido, entren en vigor automáticamente unas exigencias estrictos, mientras que si el progreso se ralentiza, estas exigencias se relajen en consecuencia. El progreso rápido e impredecible también significa que los esfuerzos de reducción de riesgos deben ser proactivos: identificar los riesgos de los sistemas de próxima generación y exigir a los desarrolladores que los aborden antes de tomar medidas de alto riesgo. Necesitamos instituciones de acción rápida y conocedoras de la tecnología para la supervisión de la IA, evaluaciones de riesgos obligatorias y mucho más rigurosas con consecuencias exigibles (incluidas evaluaciones que imponen la carga de la prueba a los desarrolladores de IA) y estándares de mitigación acordes con una IA autónoma poderosa.

Sin estos, las empresas, los ejércitos y los gobiernos pueden buscar una ventaja competitiva llevando las capacidades de IA a nuevas alturas mientras toman atajos en materia de seguridad o delegando funciones sociales clave a sistemas autónomos de IA con una supervisión humana insuficiente, cosechando los frutos del desarrollo de la IA y al mismo tiempo abandonando la sociedad. para afrontar las consecuencias. En este sentido, se proponen:

- Instituciones para gobernar la rápida evolución de IA más avanzada o de frontera: Para mantener la regulación al día frente al rápido progreso de la tecnología y evitar leyes que queden rápidamente obsoletas, las instituciones nacionales necesitan una sólida experiencia técnica y la autoridad para actuar con rapidez. Para facilitar evaluaciones y mitigaciones de riesgos técnicamente exigentes, necesitarán mucho más financiación y talento. Para abordar la actual dinámica internacional, se necesita la posibilidad de facilitar acuerdos y asociaciones internacionales. Las instituciones deben proteger el uso y la investigación académica de bajo riesgo evitando obstáculos burocráticos indebidos para modelos de IA pequeños y predecibles. El escrutinio más apremiante debería centrarse en los sistemas de IA de frontera: los pocos sistemas más potentes, entrenados en supercomputadoras de miles de millones de dólares, que tendrán las capacidades más peligrosas e impredecibles.

- Mejor conocimiento por parte de los gobiernos: Para identificar los riesgos derivados de los sistemas de IA más avanzados, los gobiernos necesitan urgentemente una visión integral del desarrollo de la IA. Los reguladores deberían exigir protección a los denunciantes, notificación de incidentes, registro de información clave sobre los sistemas de inteligencia artificial de vanguardia y los conjuntos de datos utilizados a lo

largo de su ciclo de vida, y el monitoreo del desarrollo de modelos y el uso de supercomputadoras. Los recientes desarrollos de políticas no deberían limitarse a exigir que las empresas informen los resultados de las evaluaciones de modelos voluntarios o poco especificados poco antes de su implementación. Los reguladores pueden y deben exigir que los desarrolladores de IA de vanguardia otorguen a los auditores externos un acceso in situ, integral ("caja blanca") y de ajuste fino desde el inicio del desarrollo del modelo. Esto es necesario para identificar capacidades de modelos peligrosos, como la autorreplicación autónoma, la persuasión a gran escala, la irrupción en sistemas informáticos, el desarrollo de armas (autónomas) o la ampliación del acceso a patógenos pandémicos.

- Casos de seguridad: A pesar de las evaluaciones que se realicen, no podemos considerar que los poderosos sistemas de IA de vanguardia sean "seguros a menos que se demuestre que no son seguros". Con las metodologías de prueba actuales, es fácil pasar por alto los problemas. Además, no está claro si los gobiernos podrán desarrollar rápidamente la inmensa experiencia necesaria para realizar evaluaciones técnicas confiables de las capacidades de la IA y los riesgos a escala social. Teniendo esto en cuenta, los desarrolladores de IA de vanguardia deberían asumir la carga de la prueba para demostrar que sus planes mantienen los riesgos dentro de límites aceptables. Al hacerlo, seguirían las mejores prácticas para la gestión de riesgos de industrias, como la aviación, los dispositivos médicos y el software de defensa, áreas en las que las empresas presentan casos de seguridad mediante argumentos estructurados con afirmaciones falsificables respaldadas por evidencia que identifica peligros potenciales, describe mitigaciones, muestra que los sistemas no cruzarán ciertas líneas rojas y modelan posibles resultados para evaluar el riesgo. Los casos de seguridad podrían aprovechar la profunda experiencia de los desarrolladores con sus propios sistemas. Los casos de seguridad son políticamente viables incluso cuando la gente no está de acuerdo sobre qué tan avanzada llegará a ser la IA porque es más fácil demostrar que un sistema es seguro cuando sus capacidades son limitadas. Los gobiernos no son receptores pasivos de casos de seguridad: establecen umbrales de riesgo, codifican mejores prácticas, emplean expertos y auditores externos para evaluar casos de seguridad y realizan evaluaciones de modelos independientes, y responsabilizan a los desarrolladores si sus afirmaciones de seguridad son posteriormente falsificadas.

- Mitigación: Para mantener los riesgos de la IA dentro de límites aceptables, son necesarios mecanismos de gobernanza que se ajusten a la magnitud de los riesgos. Los reguladores deben aclarar las responsabilidades legales que surgen de los marcos de responsabilidad existentes y responsabilizar legalmente a los desarrolladores y propietarios de IA de vanguardia por los daños causados por sus modelos que puedan preverse y prevenirse razonablemente, incluidos los daños que previsiblemente surjan del despliegue de potentes sistemas de IA cuyo comportamiento no pueden predecir. La responsabilidad, junto con las consiguientes evaluaciones y casos de seguridad, puede prevenir daños y crear incentivos muy necesarios para invertir en seguridad.

El documento concluye señalando que "Para orientar la IA hacia resultados positivos y alejarla de la catástrofe, debemos reorientarnos. Existe un camino responsable, si tenemos la sabiduría para seguirlo".