ANSIEDADES POR LA IA EN LOS REGÍMENES DE PROHIBICIÓN DE ARMAS QUÍMICAS Y BIOLÓGICAS

La inteligencia artificial (IA) tiene el potencial de revolucionar las actividades científicas, económicas y sociales en diversos sectores. Sin embargo, existe una creciente preocupación por el impacto negativo que podrían tener las aplicaciones de doble uso y los resultados de la IA. Estas "ansiedades por la IA" son tanto reales como imaginarias, y se basan en lo que es posible hoy y lo que podría ser posible en el futuro.

La preocupación por las posibles aplicaciones de doble uso de la IA también se ha expresado en el contexto de los regímenes de prohibición de armas químicas y biológicas (chemical and biological weapons, CBW), donde se está comenzando a considerar la capacidad de la IA para permitir el diseño, desarrollo, despliegue y detección de CBW.

Los relatos emergentes que sugieren que la IA podría facilitar el desarrollo de agentes supertóxicos, o proporcionar rutas de bajo costo para que los actores estatales y no estatales desarrollen y empleen CBW, están cautivando la imaginación del público.

Esta nota informativa presenta cuatro nuevas inquietudes por la IA en relación con las CBW, y utiliza los principios éticos actuales para una IA responsable como guía a fin de evaluar cómo nuestras normas y valores pueden ayudar a abordar estos desafíos.

Al identificar qué características de la IA buscan mitigar estos principios, podemos visualizar mejor cómo esas características de la IA podrían generar desafíos específicos dentro de los regímenes de prohibición de CBW. Este es un enfoque en etapa inicial que puede contribuir a los esfuerzos crecientes de comprender la naturaleza de estos nuevos desafíos.

Puntos clave

- La IA no es una entidad independiente, sino un componente dentro de un sistema más amplio que se combina con otras tecnologías para mejorar el procesamiento de datos y la toma de decisiones.
- Si bien la IA puede plantear nuevos desafíos y ansiedades, muchos de estos no son sustancialmente diferentes de los asociados con otras tecnologías o prácticas. La IA puede intensificar y modificar los desafíos existentes y debe considerarse en el contexto más amplio de los sistemas en los que opera.
- Mitigar los efectos potencialmente perjudiciales de la IA no requiere conjuntos completamente nuevos de arquitecturas de gobernanza, sino que, al identificar y abordar las posibles ansiedades por la IA, podría darse forma a la enmienda y mejorar los marcos existentes.

UNIVERSITY OF SUSSEX

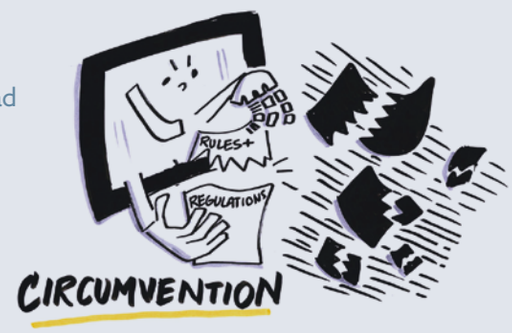
HSP | THE HARVARD SUSSEX PROGRAM

IA + (TECNOLOGÍA)

La integración de la IA con otras tecnologías plantea desafíos significativos a los regímenes de prohibición de CBW. Los escenarios que se ajustan a esta ansiedad giran en torno a las formas en que la IA podría aumentar o amplificar los riesgos de las tecnologías de doble uso existentes. Por ejemplo, el uso de objetivos de IA en vehículos autónomos para la aplicación de agentes antidisturbios (riot control agents, RCA) puede profundizar la ambigüedad en torno al empleo seguro y legal de los RCA según la Convención sobre Armas Químicas (Chemical Weapons Convention, CWC). Otra posibilidad es que la conexión de la IA con la robótica avanzada permita a los actores con recursos humanos limitados diseñar y sintetizar productos químicos o biológicos de forma remota. En este caso, los principios éticos pueden ajustarse más a los valores de 'Segura y protegida' y 'Centrada en el ser humano', donde la infraestructura, la gobernanza y las evaluaciones de intención pueden ayudar a aclarar las posibles medidas de gestión en el uso de la IA.

VÍAS A LA EVASIÓN

La IA podría ayudar a superar las limitaciones, los obstáculos y los controles tradicionales que dificultan el desarrollo, el almacenamiento y el despliegue de CBW. Por ejemplo, la IA puede tener el potencial de mejorar los esfuerzos de los actores malintencionados para eludir los controles de importación/exportación mediante la identificación o el modelado de precursores no incluidos en la lista para la síntesis química. Además, el uso de la IA puede reducir la cantidad de recursos humanos necesarios para desarrollar o desplegar estas armas, lo que reduce los costes y evita posibles 'dilemas morales'. La IA podría utilizarse también para diseñar agentes que sean más complejos de detectar y de los que puede ser más difícil defenderse. En estos casos, podríamos esperar que los valores de 'Controlabilidad' y 'Responsabilidad' reflejen la necesidad de examinar la gobernanza en el desarrollo de IA.



LA IA DEBE

Explicable y responsable

"La capacidad de explicar cómo y por qué se alcanzaron resultados particulares; y que las entradas de datos, la estructura de diseño y los sistemas operativos, la operación general y los resultados, deben permitir ejercicios de rendición de cuentas".

Controlable

"Capacidad para controlar insumos, funcionamiento y resultados, a fin de redirigir, modificar, anular o cerrar operaciones; los productos también deben tener potencial de controlabilidad (es decir, prevención de la diseminación automática)".

ERÍA SER...

Centrada en el ser humano

"Los seres humanos son, en última instancia, los responsables del diseño y el ciclo operativo completo de la IA y sus resultados, y comprender dónde y cómo están implicados los humanos puede respaldar las evaluaciones de intención".

Segura y protegida

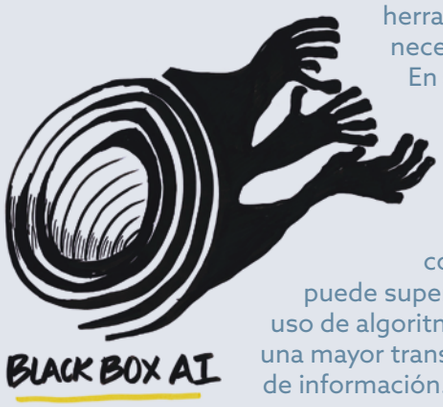
"La IA requiere seguridad cibernética y de infraestructura para protegerse de los actores malignos; internamente requiere respaldos, capacidad de detener/anular; revisar los mecanismos para garantizar que los resultados no presenten peligros para los seres humanos".

LEGITIMACIÓN PROGRESIVA

La erosión insidiosa de las prohibiciones legales y los tabúes sociales a través de ciertas líneas de investigación y desarrollo no son una preocupación nueva. Sin embargo, el potencial de la IA para facilitar, acelerar y crear nuevas vías para la investigación cuando la línea que separa lo legítimo de lo ilegítimo es muy delgada puede brindar a los malintencionados nuevas oportunidades para desarrollar agentes y tecnologías que podrían desafiar los tratados de CBW. Por ejemplo, la exacerbación de los supuestos existentes por parte de la IA puede producir un despliegue ambiguo o no intencionado de RCA que desafía los límites legales. Además, la investigación militar sobre los procesos biológicos de la vida hacia resultados de 'guerra sin muerte' podría revitalizarse con las oportunidades que ofrecen las tecnologías de IA. En este caso, es posible que los principios de 'Responsabilidad' y 'Centrada en el ser humano' proporcionen una idea de cómo los diferentes actores pueden ser gobernados (o pueden gobernar) para mantener los límites legales y sociales a la luz de las capacidades novedosas.

LA CAJA NEGRA

La naturaleza oscura de las entradas, métodos y salidas de la IA crea cargas de transparencia adicionales para desarrolladores y usuarios. Esta intangibilidad, opacidad y falta de explicabilidad de los sistemas de IA impregnarán y caracterizarán los desafíos en muchas áreas. Por ejemplo, al verificar alegaciones de uso, rastrear la rendición de cuentas puede ser complicado si se utiliza una herramienta de IA lista para usar con un conocimiento limitado de la procedencia exacta de la herramienta, los datos de entrenamiento y los métodos de deducción. Otra posibilidad es que el uso de herramientas de IA para desarrollar y producir CBW potenciales limite la atribución en los casos en que múltiples actores tengan acceso a las herramientas y los datos necesarios para hacerlo. En estos casos, los principios de 'Explicable', 'Responsable' y 'Controlable' pasan a primer plano para comprender cómo se puede supervisar la creación y el uso de algoritmos de IA a través de una mayor transparencia e intercambio de información.



Trascendencia

Las cuatro ansiedades por la IA demuestran que la integración de las tecnologías de IA no necesariamente produce categorías completamente nuevas de amenazas o riesgos, sino que se integran dentro de las ya existentes. Es importante destacar que los escenarios específicos se ajustan a múltiples categorías de ansiedad según las formas en que la IA interactúa con las tecnologías y los modos de gobernanza existentes. Este reconocimiento requiere que las inquietudes sobre la IA se basen en los contextos y desafíos con los que los actores que operan en los regímenes de prohibición de CBW ya están familiarizados. Dentro de esos contextos, los diferentes impactos de la IA probablemente serán los siguientes:

- Acelerar los procesos de investigación y desarrollo.
- Abrir nuevas vías de investigación prospectiva.
- Degradar la transparencia.
- Complicar la providencia, relevancia y consecuencia de la información.

Por lo tanto, la siguiente etapa es evaluar si las formas de gobernanza que hoy existen pueden, a través de la capacidad actual o mediante modificaciones, mitigar el impacto de la IA en los desafíos presentes e incipientes.

Los cuatro principios éticos para una IA responsable nos dirigen al tipo de preguntas que deberíamos hacernos en esos esfuerzos por visualizar, caracterizar y minimizar las implicaciones negativas de las tecnologías de IA.

Para obtener más información:

Esta investigación fue realizada por el Programa Harvard-Sussex en la Unidad de Investigación de Políticas Científicas de University of Sussex Business School. El proyecto de investigación fue financiado por el Centro de Control de Armas y No Proliferación (Counter Proliferation and Arms Control Centre) de la Oficina de Relaciones Exteriores y de la Mancomunidad de Naciones (Foreign, Commonwealth and Development Office) del Reino Unido. Nuestro agradecimiento a Shaunna McIvor y Boaz Chan, quienes colaboraron enormemente con el proyecto.

Investigador principal: **Dr. Joshua R Moon**
Correo electrónico: **J.R.Moon@sussex.ac.uk**

¿Cómo se puede gobernar mejor la IA en contextos de CBW?

1. Contextualice el desarrollo y las aplicaciones de IA para comprender mejor los desafíos específicos de las políticas.
2. Socialice la IA mediante la identificación de los actores involucrados en el desarrollo y el propósito del uso.
3. Examine las ansiedades por la IA a la luz de cómo pueden impedir o debilitar los esfuerzos actuales en la prohibición de CBW.
4. Evalúe los actores y las arquitecturas de gobernanza actuales para comprender dónde existen oportunidades y brechas para abordar estas ansiedades por la IA y, cuando sea posible, rectificarlas.
5. Concéntrese en las herramientas y los mecanismos existentes para abordar las ansiedades por la IA específicas del desafío.

Próximos pasos para una mejor política de IA:

- Construir escenarios potenciales más detallados del impacto de la IA en las prohibiciones de CBW, con datos específicos sobre el uso de las tecnologías y los roles de los diferentes actores dentro de esos escenarios, permitirá una mejor comprensión de cómo se pueden reducir los riesgos.
- Comprender los impactos y las consecuencias de la IA en el contexto y en la práctica permitirá un conocimiento más profundo de cómo se puede gestionar la IA para preservar sus impactos positivos en la sociedad.
- Delinear y evaluar la gobernanza existente para los regímenes de prohibición de CBW con referencia específica a escenarios detallados proporcionará a los formuladores de políticas una mejor comprensión de dónde interactúa la IA con diferentes redes de prevención y cómo se pueden desarrollar esas redes para adaptarse a esta nueva área de ciencia y tecnología.