Bulletin of the Atomic Scientists

Señalización costosa: cómo resaltar la intención puede ayudar a los gobiernos a evitar peligrosos errores de cálculo de la IA

By Owen J. Daniels, Andrew Imbrie | December 15, 2023

President Joe Biden hosts a meeting on artificial intelligence, Tuesday, June 20, 2023, at

The Fairmont hotel in San Francisco. Credit: The White House by Adam Schultz

En octubre, la administración Biden publicó una amplia orden ejecutiva sobre el desarrollo y uso seguro y confiable de la inteligencia artificial. El extenso texto expone la visión del presidente para el uso responsable de la IA en Estados Unidos y reconoce que "aprovechar la IA para siempre... requiere mitigar sus riesgos sustanciales". Encomienda a las agencias federales la tarea de ayudar a proteger la seguridad, la privacidad y los derechos civiles de los estadounidenses contra el uso indebido de la IA, en algunos casos con plazos a corto plazo que probablemente impulsen a la burocracia a actuar.

A pesar de tener casi 100 páginas, la orden ejecutiva no es una hoja de ruta exhaustivamente detallada para una IA confiable. Establece prioridades tecnológicas y pasos regulatorios dentro de las numerosas agencias del poder ejecutivo y describe cómo pueden comenzar a implementar salvaguardas de IA, pero no tiene el mismo poder que una ley aprobada por el Congreso. Tampoco es demasiado específico al enumerar las diversas herramientas que las agencias necesitarán para monitorear y regular la IA. No obstante, la orden ejecutiva envía una señal de fundamental importancia al público estadounidense, así como a los aliados y competidores en el extranjero, sobre la visión responsable de la IA de la administración y los pasos necesarios para hacerla realidad.

La señalización gubernamental en torno a la IA es un desafío. Los formuladores de políticas pueden tener dificultades para comunicar sus intenciones respecto de una tecnología tan transformadora, que ya está remodelando las sociedades, las economías, la diplomacia y la guerra. Con el reciente ritmo vertiginoso de los desarrollos comerciales de IA, muchos de ellos dispersos globalmente, los países pueden no estar seguros de cómo otros planean utilizar las últimas aplicaciones para obtener una ventaja competitiva. Los riesgos de una percepción errónea y una escalada involuntaria abundan, por lo que es imperativo que las autoridades aprovechen todo el conjunto de herramientas políticas para enviar señales de intención claras y creíbles.

Después de todo, hay una razón por la que Washington y Moscú establecieron una línea directa tras el casi desastre de la crisis de los misiles cubanos.

Para aclarar sus intenciones, los formuladores de políticas pueden utilizar señales costosas (una herramienta política examinada de cerca en la literatura sobre relaciones internacionales) para comunicar sobre la IA y decodificar las intenciones de otros. Las señales son costosas cuando el remitente paga un precio, ya sea político, reputacional o incluso monetario, si no cumple con los mensajes que comunica. Durante la Guerra Fría, por ejemplo, los gobiernos revelaron ciertas capacidades a sus rivales para comunicar mensajes de disuasión; Si bien tales acciones limitaron el potencial de uso sorpresa, permitieron a los adversarios comprender aspectos de nuevos sistemas revolucionarios. Aplicar el marco de señales costosas a la IA en medio del

contexto geopolítico actual puede ayudar a los formuladores de políticas a trazar un camino hacia el uso responsable de estas máquinas herramienta.

Respaldar las palabras con acciones. Discernir las intenciones de los aliados y adversarios en la IA es de vital importancia para comprender los riesgos asociados con las diferentes aplicaciones de la tecnología. Por ejemplo, a los responsables de las políticas les puede preocupar que sus homólogos de otro estado se apresuren a desplegar capacidades que no han sido probadas de forma adecuada para tener una ventaja sobre la competencia. Cuando se integran en estrategias de política exterior, defensa o tecnología, las señales costosas pueden ayudar a los formuladores de políticas a resaltar sus intenciones, mitigar los riesgos de una escalada involuntaria o una percepción errónea y revelar capacidades a los adversarios que disuaden la toma de riesgos.

A medida que Estados Unidos y China navegan por la competencia estratégica y tecnológica, la capacidad de discernir intenciones con señales costosas será clave. Sin embargo, aprovechar esta herramienta de políticas en IA probablemente será un desafío, dadas las aplicaciones de doble uso de la tecnología, el rápido progreso en grandes modelos de lenguaje y la competencia entre empresas privadas para llevar sus modelos de IA al mercado primero. Los formuladores de políticas en ambos países necesitarán monitorear los últimos avances tecnológicos y permanecer alerta a las señales que la otra parte pueda estar enviando. La elección no es simplemente "ocultar o revelar" las capacidades de la IA, sino también cómo revelarlas y a través de qué canales. El hecho de enviar un mensaje no garantiza que el receptor lo entenderá correctamente y las señales pueden perderse en medio del ruido de las grandes burocracias. La ejecución es importante, pero a veces resulta insuficiente.Cuatro tipos de señales costosas son particularmente relevantes para la IA. El primer tipo es atar las manos, lo que implica asumir compromisos públicos estratégicos ante audiencias nacionales y extranjeras. Si los países firman un tratado comprometiéndose a desarrollar y utilizar estándares responsables de IA, por ejemplo, pueden enfrentar presión de los cosignatarios y del público en caso de que implementen modelos de IA de vanguardia que no cumplan con estos estándares.

El segundo son los costos hundidos, donde el precio de un compromiso se incorpora desde el principio y el alto precio de una decisión indica que es poco probable que el remitente incumpla. En IA, podríamos pensar en los compromisos para licenciar y registrar algoritmos o en inversiones en instalaciones de infraestructura de prueba y evaluación como costos irrecuperables.

RELACIONADO:

Por qué la IA para el diseño biológico debería regularse de forma diferente a

los chatbots

El tercero son los costos a plazos, donde el remitente se compromete a sostener los costos en el futuro. Estos podrían incluir herramientas de contabilidad informática que rastrean grupos de chips de IA en centros de datos o la verificación de los compromisos gubernamentales de realizar evaluaciones de riesgo de modelos de IA y poner los resultados de esas evaluaciones a disposición del público.

El cuarto son los costos reducibles, donde el remitente paga los costos de enviar la señal por adelantado pero puede compensar esos costos con el tiempo; por ejemplo, enfoques de datos pequeños para la IA o tarjetas modelo y hojas de datos que brindan transparencia sobre los datos de entrenamiento, los pesos de los modelos y otras características específicas de los modelos de IA. Estas medidas pueden resultar costosas de implementar al principio; Sin embargo, con el tiempo, los costos pueden recuperarse a medida que los modelos ganan popularidad y las empresas que los implementan se ganan una reputación de desarrollo confiable.

El potencial y los desafíos de la señalización costosa. La decodificación de señales en torno a la IA militar proporciona un ejemplo tanto de la importancia como de los desafíos de la señalización. Esta tarea es difícil por varias razones. Las tecnologías de IA pueden fallar de maneras sorprendentes y difíciles de solucionar. Los métodos de prueba y evaluación para evaluar los sistemas militares basados en IA son incipientes, y el papel de la industria privada en el desarrollo de aplicaciones de doble uso puede alimentar percepciones y cálculos erróneos entre los Estados que los implementan.

¿Cómo pueden los países superar estos desafíos con una señalización costosa? Por un lado, los gobiernos y las empresas que desarrollan capacidades militares de IA podrían utilizar mecanismos de atar manos, asumiendo compromisos públicos para comunicar la intención sobre dónde usarán o no la IA. La “Declaración política sobre el uso militar responsable de la inteligencia artificial y la autonomía” del Departamento de Estado de EE. UU., publicada en febrero y que cuenta con más de 40 signatarios, consagra el compromiso de “garantizar que la seguridad y la eficacia de las capacidades militares de IA estén sujetas a pruebas y garantías apropiadas y rigurosas dentro de sus usos bien definidos y durante todo su ciclo de vida”. Aunque Estados Unidos podría retractarse de este compromiso, incurriría en costos políticos y de reputación entre aliados y competidores. De manera similar, China podría utilizar una señal de atar las manos, como permitir que el Ejército Popular de Liberación discuta medidas de reducción del riesgo de IA con Estados Unidos (una opción que supuestamente rechazó durante las conversaciones bilaterales de defensa en 2021) para indicar que se toma en serio la seguridad internacional de la IA. normas y reducir los riesgos crecientes.

Además, las medidas de costos hundidos para reducir los riesgos podrían incluir invertir en infraestructura de prueba y evaluación para la IA militar y aumentar la transparencia en torno a las mejores prácticas de seguridad. Compartir información podría ayudar a todas las partes a comprender mejor dónde se emplean los sistemas basados en IA en la toma de decisiones militares y políticas. Por ejemplo, si China integrara la IA en los sistemas de alerta temprana (sistemas cuyo fallo sirvió como fuente de aterradores casi accidentes durante la Guerra Fría), ¿los líderes chinos en una crisis actual considerarían esos fallos como percances no intencionados o preludios de un ataque intencional? Dadas las incertidumbres en torno a las leyes, doctrinas y políticas pertinentes para gestionar incidentes relacionados con sistemas habilitados para IA, una crisis que involucre tales plataformas fácilmente podría convertirse en un conflicto.

Estados Unidos, China y otras naciones podrían agregar costos de pago a plazos a los costos irrecuperables comprometiéndose públicamente a compartir información, medidas de transparencia e inspecciones de modelos de IA, sitios de prueba y hardware designados en un "Acuerdo Internacional de Incidentes Autónomos". Los compromisos públicos, junto con los costos de puesta en marcha y mantenimiento de las inversiones en infraestructura de prueba, darían credibilidad a las promesas de cumplir con las normas internacionales en IA militar.

Los gobiernos podrían utilizar costos reducibles para medidas de fomento de la confianza en torno a la IA militar, financiando o asociándose con la industria y el mundo académico para apoyar la investigación de IA interpretable, creando incentivos para mejorar la transparencia del desarrollo de modelos de IA a través de tarjetas modelo o ejercicios de equipos rojos. Invertir en entidades de investigación globales que puedan monitorear y medir las capacidades de IA o mejorar.

Por ejemplo, dos décadas de guerra y operaciones antiterroristas en el Medio Oriente consolidaron a las monarquías del Golfo como socios regionales de Estados Unidos. Sin embargo, recientemente, los responsables políticos estadounidenses han planteado preocupaciones de seguridad por el estrechamiento de los vínculos entre China y el Golfo, incluso en inteligencia artificial y en infraestructura de telecomunicaciones 5G. Las señales sobre la IA democrática podrían llevar a estados autoritarios como los del Golfo a preferir capacidades de IA fabricadas en China, dependiendo de cómo interpreten las señales de la IA democrática, fortaleciendo potencialmente la influencia de China. Escenarios similares podrían surgir en otros teatros.

Estados Unidos no debería amortiguar sus costosas señales en torno a la IA democrática basándose en sus relaciones con dichos estados. Sin embargo, los diplomáticos y estrategas estadounidenses deben ser conscientes de las implicaciones posteriores que podrían tener las costosas señales democráticas

de IA y los consiguientes desafíos diplomáticos.

Señales para una nueva era. Las crisis impulsadas por percepciones erróneas no son nuevas en las relaciones internacionales, pero las aplicaciones multipropósito de la IA, el enredo del sector privado y la proliferación más allá de los gobiernos significan que las señales actuales no son necesariamente "altas y claras" en comparación con eras anteriores en el arte de gobernar diplomático. Las señales pueden resultar inadvertidamente costosas a la hora de llegar a diferentes audiencias, y para que sean verdaderamente efectivas deben integrarse en estrategias integrales que incorporen diferentes instrumentos de política. En el entorno de información competitivo y multifacético actual, hay aún más actores con influencia en el panorama de la señalización. El contexto es clave para transmitir señales de forma clara y creíble.

Un camino a seguir es que los gobiernos aprovechen las prácticas y regulaciones de adquisiciones para dar forma a las normas en torno al desarrollo y uso de la IA. Por ejemplo, los formuladores de políticas podrían trabajar con expertos de la industria e investigadores académicos para consagrar normas en torno a la transparencia de la IA (como la publicación de tarjetas modelo, tarjetas de sistema o documentación similar) a través de políticas de adquisiciones, incluidas protecciones adecuadas para la privacidad y la seguridad. Los formuladores de políticas también deberían considerar la posibilidad de incorporar señales costosas en los diálogos y ejercicios teóricos con aliados y competidores para aclarar suposiciones, mitigar los riesgos crecientes y desarrollar entendimientos compartidos en torno a las comunicaciones de crisis.

En cuanto a la orden ejecutiva de la administración, es difícil saber con certeza si el equipo del presidente Biden tenía la intención de enviar señales costosas en la forma precisa que se describe aquí. No todas las señales son intencionales, y los actores comerciales pueden calcular los costos de manera diferente a los gobiernos o a los actores de la industria en otros países. No obstante, se puede considerar que la orden ejecutiva ata las manos al señalar públicamente el compromiso de la administración con la IA responsable, además de emplear una combinación de señales costosas al pedir medidas a las agencias federales, como el nombramiento de directores generales de IA, la colocación de marcas de agua en las comunicaciones oficiales, la evaluación y la racionalización. criterios de visa para traer inmigrantes talentosos a los Estados Unidos y establecer salvaguardias en áreas como la bioseguridad. La orden ejecutiva bien puede indicar a los competidores, como China, o a los aliados en Europa, que están desarrollando sus propios estándares y regulaciones de IA, que el gobierno de Estados Unidos está evaluando seriamente la mejor manera de implementar y capitalizar la aplicación responsable de la IA.

Que los aliados y competidores reciban estas señales como se esperaba es otra historia. Se espera que no sea necesaria otra crisis de los misiles cubanos para que los países que despliegan IA establezcan líneas abiertas de comunicación y eviten una escalada impulsada por las nuevas tecnologías. Las señales pueden ser ruidosas y en ocasiones confundir a algunas audiencias, pero aun así son necesarias.

Las opiniones y caracterizaciones de este artículo son de los autores y no necesariamente representan las del gobierno de Estados Unidos.