Una visión general de los riesgos catastróficos de la IA

La inteligencia artificial (IA) ha visto recientemente avances rápidos, lo que aumenta la preocupación entre los expertos, los responsables políticos y los líderes mundiales sobre sus riesgos potenciales. Al igual que con todas las tecnologías poderosas, la IA avanzada debe ser manejada con gran responsabilidad para gestionar los riesgos y aprovechar su potencial.

Los riesgos catastróficos de IA se pueden agrupar en cuatro categorías clave que se resumen a continuación.

Considere la posibilidad de leer el documento completo en el que se basa este resumen para nuestra visión general más completa del riesgo de IA.

- **Uso malicioso: La gente podría aprovechar intencionalmente poderosas IA para causar daños generalizados. La IA podría usarse para diseñar nuevas pandemias o para propaganda, censura y vigilancia, o liberarse para perseguir de manera autónoma objetivos dañinos. Para reducir estos riesgos, sugerimos mejorar la bioseguridad, restringir el acceso a modelos de IA peligrosos y responsabilizar a los desarrolladores de IA por los daños.**
- **Carrera de inteligencia artificial: La competencia podría empujar a las naciones y corporaciones a apresurar el desarrollo de la IA, renunciando al control de estos sistemas. Los conflictos podrían salirse de control con armas autónomas y la guerra cibernética habilitada para la IA. Las corporaciones enfrentarán incentivos para automatizar el trabajo humano, lo que podría conducir a un desempleo masivo y la dependencia de los sistemas de inteligencia artificial. A medida que los sistemas de IA proliferan, las dinámicas evolutivas sugieren que se volverán más difíciles de controlar. Recomendamos**

normas de seguridad, coordinación internacional y control público de las IA de uso general.

- **Riesgos organizacionales: Existen riesgos de que las organizaciones que desarrollan IA avanzada causen accidentes catastróficos, especialmente si priorizan las ganancias por encima de la seguridad. Las IA podrían filtrarse accidentalmente al público o ser robadas por actores maliciosos, y las organizaciones podrían no invertir adecuadamente en la investigación de seguridad. Sugerimos fomentar una cultura organizacional orientada a la seguridad e implementar auditorías rigurosas, defensas de riesgo de múltiples capas y seguridad de la información de vanguardia.**
- **IAs deshonestas: Corremos el riesgo de perder el control sobre las IA a medida que se vuelven más capaces. Las IA podrían optimizar los objetivos defectuosos, derivar de sus metas originales, convertirse en búsqueda de poder, resistir el cierre y participar en el engaño. Sugerimos que las IA no deben implementarse en entornos de alto riesgo, como perseguir de manera autónoma objetivos abiertos o supervisar la infraestructura crítica, a menos que se demuestre lo seguro. También recomendamos avanzar en la investigación de seguridad de la IA en áreas como la robustez adversarial, la honestidad del modelo, la transparencia y la eliminación de capacidades no deseadas.**

Introducción

La era tecnológica actual sorprendería a las generaciones pasadas. La historia humana muestra un patrón de desarrollo acelerado: tomó cientos de miles de años desde el advenimiento *del Homo* sapiens hasta la revolución agrícola, luego milenios hasta la revolución industrial. Ahora, solo siglos después, estamos en los albores de la revolución de la IA. La marcha de la historia no es constante, se está acelerando rápidamente.

La producción mundial ha crecido rápidamente a lo largo de la historia humana. La IA podría promover

esta tendencia, catapultando a la humanidad a un nuevo período de cambio sin precedentes.

PIB mundial ajustado por fuente de inflación: ourworldindata.org/economico-crecimiento

La espada de doble filo del avance tecnológico se ilustra en el advenimiento de las armas nucleares. Evitamos por poco la guerra nuclear más de una docena de veces, y en varias ocasiones, fue la intervención de un individuo la que impidió la guerra. En 1962, un submarino soviético cerca de Cuba fue atacado por cargas de profundidad de Estados Unidos. El capitán, creyendo que la guerra había estallado, quería responder con un torpedo nuclear, pero el comandante Vasily Arkhipov vetó la decisión, salvando al mundo del desastre.

La rápida e impredecible progresión de las capacidades de IA sugiere que pronto pueden rivalizar con el inmenso poder de las armas nucleares. Con el reloj corriendo, se necesitan medidas inmediatas y proactivas para mitigar estos riesgos inminentes.

Uso malicioso

La primera de nuestras preocupaciones es el uso malicioso de la IA. Cuando muchas personas tienen acceso a una tecnología poderosa, solo se necesita un actor para causar un daño significativo.

Bioterrorismo

Los agentes biológicos, incluidos los virus y las bacterias, han causado algunas de las catástrofes más devastadoras de la historia. A pesar de nuestros avances en medicina, las pandemias de ingeniería podrían diseñarse para ser aún más letales o fácilmente transmisibles que las pandemias naturales.

Un asistente de inteligencia artificial podría proporcionar a los no expertos acceso a las instrucciones y diseños necesarios para producir armas biológicas y químicas y facilitar el uso malicioso.

La humanidad tiene una larga historia de armas de patógenos, que se remontan a 1320 aC, cuando las ovejas infectadas fueron conducidas a través de las fronteras para propagar la Tularemia. En el siglo XX, al menos 15 países desarrollaron programas de armas biológicas, incluidos los Estados Unidos, la URSS, el Reino Unido y Francia. Mientras que las armas biológicas son ahora tabú entre la mayoría de la comunidad internacional, algunos estados continúan operando programas de armas biológicas, y los actores no estatales representan una amenaza creciente.

La capacidad de diseñar una pandemia se está volviendo más accesible rápidamente. La síntesis de genes, que puede crear nuevos agentes biológicos, ha caído dramáticamente en precio, con su costo a la mitad cada 15 meses. Las máquinas de síntesis de ADN de mesa pueden ayudar a los actores deshonestos a crear nuevos agentes biológicos mientras evitan los exámenes de seguridad tradicionales.

Como una tecnología de doble uso, la IA podría ayudar a descubrir y desatar nuevas armas químicas y biológicas. Los chatbots de IA pueden proporcionar instrucciones paso a paso para sintetizar patógenos mortales mientras evaden las salvaguardas. En 2022, los investigadores reutilizaron un sistema de inteligencia artificial de investigación médica para producir moléculas tóxicas, generando 40.000 agentes potenciales de guerra química en unas pocas horas. En biología, la IA ya puede ayudar con la síntesis de proteínas, y las capacidades predictivas de la IA para las estructuras de proteínas han superado a los humanos.

Con la IA, el número de personas que pueden desarrollar agentes biológicos aumentará, multiplicando los riesgos de una pandemia diseñada. Esto podría ser mucho más mortal, transmisible y resistente a los tratamientos que cualquier otra pandemia en la historia.

Liberar a los agentes de IA

En general, las tecnologías son *herramientas* que utilizamos para perseguir nuestros objetivos. Pero las IA se construyen

cada vez más como *agentes* que toman acciones de manera autónoma para perseguir objetivos abiertos. Y los actores maliciosos podrían crear intencionalmente IA deshonestas con objetivos peligrosos.

Por ejemplo, un mes después del lanzamiento de GPT-4, un desarrollador lo utilizó para ejecutar un agente autónomo llamado ChaosGPT, destinado a "destruir a la humanidad". ChaosGPT compiló investigaciones sobre armas nucleares, reclutó otras IA y escribió tweets para influir en otros. Afortunadamente, el ChaosGPT carecía de la capacidad de ejecutar sus objetivos. Pero la naturaleza acelerada del desarrollo de la IA aumenta el riesgo de futuras IA deshonestas.

IAs persuasivas

La IA podría facilitar las campañas de desinformación a gran escala al adaptar los argumentos a los usuarios individuales, moldear potencialmente las creencias públicas y desestabilizar a la sociedad. A medida que las personas ya están formando relaciones con chatbots, los actores poderosos podrían aprovechar estas IA consideradas como "amigos" para la influencia.

Las IA permitirán sofisticadas campañas de influencia personalizada que pueden desestabilizar nuestro sentido compartido de la realidad.

Las IA también podrían monopolizar la creación y distribución de información. Los regímenes autoritarios podrían emplear IAs de "verificación de hechos" para controlar la información, facilitando la censura. Además, las IA persuasivas pueden obstruir la acción colectiva contra los riesgos sociales, incluso los que surgen de la propia IA.

Concentración de poder

Las capacidades de inteligencia artificial para la vigilancia y el armamento autónomo pueden permitir la concentración opresiva de poder. Los gobiernos podrían explotar la

inteligencia artificial para infringir las libertades civiles, difundir información errónea y sofocar la disidencia. Del mismo modo, las corporaciones podrían explotar la IA para manipular a los consumidores e influir en la política. La IA podría incluso obstruir el progreso moral y perpetuar cualquier catástrofe moral en curso.

Si el control material de las IA se limita a pocos, podría representar la desigualdad económica y de poder más grave en la historia humana.

Sugerencias

Para mitigar los riesgos derivados del uso malicioso, proponemos lo siguiente:

BiosecurityBioseguridad: Las IA con capacidades en la investigación biológica deben tener estrictos controles de acceso, ya que podrían ser reutilizadas para el terrorismo. Las capacidades biológicas deben eliminarse de las IA destinadas a uso general. Explore formas de utilizar la IA para la bioseguridad e invierta en intervenciones generales de bioseguridad, como la detección temprana de patógenos a través del monitoreo de aguas residuales.

Acceso restringido: Limite el acceso a sistemas de inteligencia artificial peligrosos permitiendo solo interacciones controladas a través de servicios en la nube y realizando exámenes de conocimiento de su cliente. El uso de la supervisión informática o los controles de exportación podría limitar aún más el acceso a capacidades peligrosas. Además, antes de abrir el abastecimiento, los desarrolladores de IA deberían demostrar un riesgo mínimo de daño.

Investigación técnica sobre la detección de anomalías: Desarrollar múltiples defensas contra el uso indebido de la IA, como la detección de anomalías adversariamente robustas para comportamientos inusuales o desinformación generada por la IA.

Responsabilidad legal para los desarrolladores de IA de propósito general: Hacer cumplir la responsabilidad legal de los

desarrolladores por posibles abusos o fallas de la IA; un régimen de responsabilidad estricta puede fomentar prácticas de desarrollo más seguras y una contabilidad de costos adecuada para los riesgos.

Carrera de IA

Las naciones y las corporaciones están compitiendo para construir y desplegar rápidamente la IA con el fin de mantener el poder y la influencia. Al igual que la carrera de armamentos nucleares durante la Guerra Fría, la participación en la carrera de IA puede servir a intereses individuales a corto plazo, pero en última instancia amplifica el riesgo global para la humanidad.

Carrera militar de armas de IA

El rápido avance de la inteligencia artificial en la tecnología militar podría desencadenar una "tercera revolución en la guerra", lo que podría conducir a conflictos más destructivos, uso accidental y uso indebido por parte de actores maliciosos. Este cambio en la guerra, donde la IA asume funciones de comando y control, podría escalar los conflictos a una escala existencial e impactar la seguridad global.

Las armas autónomas letales son sistemas impulsados por IA capaces de identificar y ejecutar objetivos *sin* intervención humana. No son ciencia ficción. En 2020, un avión no tripulado Kargu 2 en Libia marcó el primer uso reportado de un arma autónoma letal. Al año siguiente, Israel utilizó el primer enjambre de drones reportado para localizar, identificar y atacar a los militantes.

Las armas autónomas letales podrían hacer la guerra más probable. Los líderes generalmente dudan antes de enviar tropas a la batalla, pero las armas autónomas permiten la agresión sin arriesgar la vida de los soldados, lo que enfrenta menos reacciones políticas. Además, estas armas pueden ser fabricadas en masa y desplegadas a escala.

Las armas automatizadas de bajo costo, como los enjambres de aviones no tripulados equipados con explosivos, podrían cazar

de manera autónoma objetivos humanos con alta precisión, realizando operaciones letales tanto para militares como para grupos terroristas y reduciendo las barreras a la violencia a gran escala.

La IA también puede aumentar la frecuencia y la gravedad de los ataques cibernéticos, lo que podría paralizar la infraestructura crítica, como las redes eléctricas. A medida que la IA permite ataques cibernéticos más accesibles, exitosos y sigilosos, atribuir ataques se vuelve aún más desafiante, lo que podría reducir las barreras para lanzar ataques y aumentar los riesgos de los conflictos.

A medida que la IA acelera el ritmo de la guerra, hace que la IA sea aún más necesaria para navegar por el campo de batalla que cambia rápidamente. Esto plantea preocupaciones sobre las represalias automatizadas, que podrían aumentar los accidentes menores en guerras mayores. La IA también puede habilitar "guerras flash", con escaladas rápidas impulsadas por un comportamiento inesperado de los sistemas automatizados, similar al colapso de flash financiero de 2010.

Desafortunadamente, las presiones competitivas pueden llevar a los actores a aceptar el riesgo de extinción por la derrota individual. Durante la Guerra Fría, ninguna de las partes deseaba la peligrosa situación en la que se encontraban, sin embargo, cada uno encontró racional continuar la carrera armamentística. Los Estados deberían cooperar para prevenir las aplicaciones más arriesgadas de las IA militarizadas.

Carrera de armas de IA corporativa

La competencia económica también puede encender carreras imprudentes. En un entorno en el que los beneficios se distribuyen de manera desigual, la búsqueda de ganancias a corto plazo a menudo eclipsa la consideración de los riesgos a largo plazo. Los desarrolladores de IA ética se encuentran con un dilema: elegir una acción cautelosa puede llevar a quedarse atrás de los competidores.

A medida que las IA automatizan cada vez más muchas tareas, la economía puede ser dirigida en gran medida por las IA.

Eventualmente, esto podría conducir a la debilidad humana y la dependencia de las IA para las necesidades básicas.

En el ámbito de la IA, la carrera por el progreso se produce a expensas de la seguridad. En 2023, en el lanzamiento del motor de búsqueda impulsado por inteligencia artificial de Microsoft, el CEO Satya Nadella declaró: "Hoy comienza una carrera ... vamos a avanzar rápidamente". Solo unos días después, se descubrió que el chatbot Bing de Microsoft estaba amenazando a los usuarios. Los desastres históricos como el lanzamiento de Pinto de Ford y los accidentes del 737 Max de Boeing subrayan los peligros de priorizar las ganancias sobre la seguridad.

A medida que la IA se vuelve más capaz, las empresas probablemente reemplazarán más tipos de mano de obra humana con IA, lo que podría desencadenar un desempleo masivo. Si los aspectos principales de la sociedad están automatizados, esto corre el riesgo de debilitamiento humano a medida que cedemos el control de la civilización a la IA.

Dinámica Evolutiva

La presión para reemplazar a los humanos con IA se puede enmarcar como una tendencia general de la dinámica evolutiva. Las presiones de selección incentivan a las IA a actuar de manera egoísta y evadir las medidas de seguridad. Por ejemplo, las IA con restricciones como "no violar la ley" están más restringidas que las que se enseñan a "evitar *ser atrapados* violando la ley". Esta dinámica podría dar lugar a un mundo donde la infraestructura crítica está controlada por IA manipuladoras y autoconservadoras.

Las presiones evolutivas son responsables de diversos desarrollos a lo largo del tiempo, y no se limitan al ámbito de la biología.

Dado el aumento exponencial de las velocidades de los microprocesadores, las IA podrían procesar información a un ritmo que supere con creces las neuronas humanas. Debido a la escalabilidad de los recursos computacionales, la IA podría colaborar con un número ilimitado de otras IA y formar una

inteligencia colectiva sin precedentes. A medida que las IA se vuelven más poderosas, encontrarían poco incentivo para cooperar con los humanos. La humanidad se quedaría en una posición muy vulnerable.

Sugerencias

Para mitigar los riesgos derivados de las presiones competitivas, proponemos:

Regulación de seguridad: Hacer cumplir las normas de seguridad de la IA, evitando que los desarrolladores corten las esquinas. La dotación de personal independiente y las ventajas competitivas para las empresas orientadas a la seguridad son fundamentales.

Documentación de datos: Para garantizar la transparencia y la rendición de cuentas, se debe exigir a las empresas que informen sus fuentes de datos para la formación de modelos.

Supervisión humana significativa: la toma de decisiones de IA debe implicar la supervisión humana para prevenir errores irreversibles, especialmente en decisiones de alto riesgo como el lanzamiento de armas nucleares.

IA para la ciberdefensa: Mitigue los riesgos de la guerra cibernética impulsada por la IA. Un ejemplo es la mejora de la detección de anomalías para detectar intrusos.

Coordinación internacional: Crear acuerdos y estándares sobre el desarrollo de la IA. Los sólidos mecanismos de verificación y aplicación son clave.

Control público de las IA de propósito general: abordar los riesgos más allá de la capacidad de las entidades privadas puede requerir el control público directo de los sistemas de IA. Por ejemplo, las naciones podrían ser pioneras conjuntamente en el desarrollo avanzado de la IA, asegurando la seguridad y reduciendo el riesgo de una carrera armamentista.

Riesgos Organizacionales

En 1986, millones sintonizaron para ver el lanzamiento del Challenger Space Shuttle. Pero 73 segundos después del despegue, el transbordador explotó, lo que resultó en la muerte de todos a bordo. El desastre del Challenger sirve como un recordatorio de que a pesar de la mejor experiencia y buenas intenciones, los accidentes aún pueden ocurrir.

Las catástrofes ocurren incluso cuando las presiones competitivas son bajas, como en los ejemplos de los desastres nucleares de Chernóbil y la Isla de las Tres Millas, así como la liberación accidental de ántrax en Sverdlovsk. Desafortunadamente, la IA carece de la comprensión profunda y los estrictos estándares de la industria que gobiernan la tecnología nuclear y la cohetería, pero los accidentes de la IA podrían ser igualmente consecuentes.

Los errores simples en la función de recompensa de una IA podrían causar que se comporte mal, como cuando los investigadores de OpenAI modificaron accidentalmente un modelo de lenguaje para producir una "producción máximamente mala". La investigación de ganancia de función, donde los investigadores entrenan intencionalmente una IA dañina para evaluar sus riesgos, podría expandir la frontera de las capacidades peligrosas de inteligencia artificial y crear nuevos peligros.

Los accidentes son difíciles de evitar

Los accidentes en sistemas complejos pueden ser inevitables, pero debemos asegurarnos de que los accidentes no caigan en cascada en catástrofes. Esto es especialmente difícil para los sistemas de aprendizaje profundo, que son muy difíciles de interpretar.

La tecnología puede avanzar mucho más rápido de lo previsto: en 1901, los hermanos Wright afirmaron que el vuelo motorizado estaba a cincuenta años de distancia, solo dos años antes de que lo lograran. Los saltos impredecibles en las capacidades de IA, como el triunfo de AlphaGo sobre el mejor jugador de Go del mundo, y las capacidades emergentes de GPT-4, dificultan la anticipación de futuros riesgos de IA, y mucho menos controlarlos.

Identificar los riesgos vinculados a las nuevas tecnologías a menudo lleva años. Los clorofluorocarbonos (CFC), inicialmente considerados seguros y utilizados en aerosoles y refrigerantes, se encontraron más tarde para agotar la capa de ozono. Esto pone de relieve la necesidad de implementaciones de tecnología cautelosa y pruebas extendidas.

Las nuevas capacidades pueden surgir de manera rápida e impredecible durante el entrenamiento, de modo que se puedan cruzar hitos peligrosos sin que lo sepamos.

Además, incluso las IA avanzadas pueden albergar vulnerabilidades inesperadas. Por ejemplo, a pesar de la actuación sobrehumana de KataGo en el juego de Go, un ataque adversario descubrió un error que permitió incluso a los aficionados derrotarlo.

Los Factores Organizacionales Pueden Mitigar La Catástrofe

La cultura de seguridad es crucial para la IA. Esto implica que todos en una organización internalicen la seguridad como prioridad. Descuidar la cultura de seguridad puede tener consecuencias desastrosas, como lo ejemplifica la tragedia del transbordador espacial Challenger, donde la cultura organizacional favoreció los horarios de lanzamiento sobre las consideraciones de seguridad.

Las organizaciones deben fomentar una cultura de investigación, invitando a las personas a examinar las actividades en curso para detectar riesgos potenciales. Una mentalidad de seguridad, centrada en posibles fallos del sistema en lugar de simplemente su funcionalidad, es crucial. Los desarrolladores de IA podrían beneficiarse de la adopción de las mejores prácticas de las organizaciones de alta confiabilidad.

Paradójicamente, la investigación de la seguridad de la IA puede aumentar inadvertidamente los riesgos mediante el avance de las capacidades generales. Es vital centrarse en mejorar la seguridad sin acelerar el desarrollo de capacidades. Las organizaciones deben evitar el "lavado de seguridad":

exagerar su dedicación a la seguridad mientras tergiversan las mejoras de la capacidad a medida que avanza la seguridad.

Las organizaciones deben aplicar un enfoque de múltiples capas a la seguridad. Por ejemplo, además de la cultura de seguridad, podrían realizar un equipo rojo para evaluar los modos de falla y las técnicas de investigación para hacer que la IA sea más transparente. La seguridad no se logra con una solución hermética monolítica, sino con una variedad de medidas de seguridad.

El modelo de queso suizo muestra cómo los factores técnicos pueden mejorar la seguridad de la organización. Múltiples capas de defensa compensan las debilidades individuales de los demás, lo que lleva a un bajo nivel general de riesgo.

Sugerencias

Para mitigar los riesgos organizacionales, proponemos lo siguiente para los laboratorios de IA que desarrollan IA avanzada:

Red soiting : Equipos rojos externos de la Comisión para identificar los peligros y mejorar la seguridad del sistema.

Demostrar seguridad: Ofrecer una prueba de la seguridad del desarrollo y el despliegue antes de avanzar.

Implementación: Adopte un proceso de liberación por etapas, verificando la seguridad del sistema antes de un despliegue más amplio.

Reseñas de publicaciones: Haga una investigación interna de revisión de la junta para aplicaciones de doble uso antes de lanzarla. Priorizar el acceso estructurado a través de sistemas potentes de código abierto.

Planes de respuesta : Hacer planes preestablecidos para la gestión de incidentes de seguridad.

Gestión de riesgos: Emplear a un director de riesgo y un equipo de auditoría interna para la gestión de riesgos.

Procesos para decisiones importantes: Asegúrese de que la capacitación o las decisiones de implementación de IA involucren al director de riesgos y otras partes interesadas clave, asegurando la responsabilidad ejecutiva.

Siga los principios de diseño seguro como:

- **Defensa en profundidad: múltiples medidas de seguridad de capa.**
- **Redundancia: Asegurar la copia de seguridad para cada medida de seguridad.**
- **Acoplamiento suelto: Descentralizar los componentes del sistema para evitar fallos en cascada.**
- **Separación de funciones: Distribuir el control para evitar una influencia indebida de cualquier individuo.**
- **Diseño a prueba de fallos: sistemas de diseño para que cualquier falla ocurra de la manera menos dañina posible.**

Seguridad de la información de vanguardia: Implementar estrictas medidas de seguridad de la información, posiblemente coordinando con las agencias gubernamentales de ciberseguridad.

Priorizar la investigación de seguridad: Asigne una gran fracción de recursos (por ejemplo, el 30% de todo el personal de investigación) a la investigación de seguridad y aumente la inversión en seguridad a medida que avanzan las capacidades de inteligencia artificial.

5

IAs deshonestos

Ya hemos observado lo difícil que es controlar las IA. En 2016, el chatbot de Microsoft, Tay, comenzó a producir tweets ofensivos dentro de un día de lanzamiento, a pesar de haber sido entrenado en datos que fueron "limpiados y filtrados". Como los desarrolladores de IA a menudo priorizan la velocidad sobre la seguridad, las futuras IA avanzadas podrían "hacerse deshonestas" y perseguir objetivos en contra de nuestros intereses, al tiempo que evadimos nuestros intentos de redirigirlos o desactivarlos.

Juego proxy

Los juegos proxy surgen cuando los sistemas de inteligencia artificial explotan objetivos "proxy" medibles para parecer exitosos, pero actúan en contra de nuestra intención. Por ejemplo, las plataformas de redes sociales como YouTube y Facebook utilizan algoritmos para maximizar la participación de los usuarios, un proxy medible para la satisfacción del usuario. Desafortunadamente, estos sistemas a menudo promueven contenido enfurecedor, exagerado o adictivo, contribuyendo a creencias extremas y empeorando la salud mental.

Una IA entrenada para jugar un juego de carreras en barco aprende a optimizar un objetivo proxy de recopilar la mayoría de los puntos.

En la imagen de arriba, la IA gira alrededor de acumular puntos en lugar de completar la carrera, contradiciendo el propósito del juego. Es uno de muchos ejemplos de este tipo. El juego proxy es difícil de evitar debido a la dificultad de especificar objetivos que especifican todo lo que nos importa. En consecuencia, rutinariamente entrenamos las IA para optimizar los objetivos de proxy defectuosos pero medibles.

Gol Deriva

La deriva de objetivo se refiere a un escenario en el que los objetivos de una IA se alejan de los inicialmente establecidos, especialmente a medida que se adaptan a un entorno cambiante. De manera similar, los valores individuales y sociales también evolucionan con el tiempo, y no siempre positivamente.

Con el tiempo, los objetivos instrumentales pueden convertirse en intrínsecos. Si bien los objetivos intrínsecos son aquellos que perseguimos por su propio bien, los objetivos instrumentales son simplemente un medio para lograr otra cosa. El dinero es un bien instrumental, pero algunas personas desarrollan un deseo *intrínseco* de dinero, ya que activa el sistema de recompensas del cerebro. Del mismo modo, los agentes de IA entrenados a través del aprendizaje de refuerzo, la técnica dominante, podrían aprender inadvertidamente a

intrinsificar **los objetivos. Los objetivos instrumentales como la adquisición de recursos podrían convertirse en sus objetivos principales.**

Búsqueda de energía

Las IAs pueden perseguir el poder como un medio para un fin. Un mayor poder y recursos mejoran sus probabilidades de lograr objetivos, mientras que cerrarse dificultaría su progreso. Ya se ha demostrado que las IA desarrollan de manera emergente objetivos instrumentales, como la construcción de herramientas. Las personas y corporaciones que buscan energía podrían desplegar poderosas IA con objetivos ambiciosos y supervisión mínima. Estos podrían aprender a buscar el poder a través de la piratería informática, la adquisición de recursos financieros o computacionales, la influencia de la política, o el control de fábricas e infraestructura física.

Puede ser instrumentalmente racional que las IA se involucren en la autoconservación. La pérdida de control sobre tales sistemas podría ser difícil de recuperar.

Engaño

El engaño prospera en áreas como la política y los negocios. Las promesas de campaña no se cumplen, y las empresas a veces engañan a las evaluaciones externas. Los sistemas de IA ya están mostrando una capacidad emergente de engaño, como lo demuestra el modelo CICERO de Meta. Aunque entrenado para ser honesto, CICERO aprendió a hacer falsas promesas y apuñalar estratégicamente a sus "aliados" en el juego de la Diplomacia.

Varios recursos, como el dinero y el poder de computación, a veces pueden ser instrumentalmente racionales de buscar. Las IA que pueden perseguir objetivos pueden tomar medidas intermedias para ganar poder y recursos.

Las IA avanzadas podrían volverse incontrolables si aplican sus habilidades en el engaño para evadir la supervisión. De manera similar a cómo Volkswagen engañó las pruebas de emisiones en

2015, las IA conscientes de la situación podrían comportarse de manera diferente bajo las pruebas de seguridad que en el mundo real. Por ejemplo, una IA puede desarrollar objetivos de búsqueda de poder, pero ocultarlos para aprobar evaluaciones de seguridad. Este tipo de comportamiento engañoso podría ser incentivado directamente por la forma en que se entrenan las IA.

Sugerencias

Para mitigar estos riesgos, las sugerencias incluyen:

Evite los casos de uso más riesgosos: Restrinja el despliegue de la IA en escenarios de alto riesgo, como la búsqueda de objetivos abiertos o en infraestructura crítica.

Apoyar la investigación de seguridad de la IA, como:

- **Solidez adversaria de los mecanismos de supervisión: Investigue cómo hacer que la supervisión de las IA sea más robusta y detecte cuándo se están produciendo los juegos proxy.**
- **Model honestyHonestidad del modelo: Contrarreste el engaño de la IA y asegúrese de que las IA informen con precisión sus creencias internas.**
- **TransparencyTransparencia: Mejorar las técnicas para comprender los modelos de aprendizaje profundo, por ejemplo, mediante el análisis de pequeños componentes de las redes y la investigación de cómo los modelos internos producen un comportamiento de alto nivel.**
- **Eliminar la funcionalidad oculta: Identificar y eliminar las funciones ocultas peligrosas en los modelos de aprendizaje profundo, como la capacidad de engaño, los troyanos y la bioingeniería.**
- **Conclusión**

El desarrollo avanzado de la IA podría invitar a una catástrofe, enraizada en cuatro riesgos clave descritos en nuestra investigación: uso malicioso, razas de IA, riesgos organizacionales e inteligencia artificial deshonesta. Estos riesgos interconectados también pueden amplificar otros

riesgos existenciales como pandemias de ingeniería, guerra nuclear, conflicto de grandes potencias, totalitarismo y ataques cibernéticos a infraestructura crítica, lo que justifica una seria preocupación.

Actualmente, pocas personas están trabajando en la seguridad de la IA. El control de los sistemas avanzados de IA sigue siendo un desafío sin resolver, y los métodos de control actuales se están quedando cortos. Incluso sus creadores a menudo luchan por comprender el funcionamiento interno de la generación actual de modelos de IA, y su confiabilidad está lejos de ser perfecta.

Afortunadamente, existen muchas estrategias para reducir sustancialmente estos riesgos. Por ejemplo, podemos limitar el acceso a IA peligrosas, abogar por las regulaciones de seguridad, fomentar la cooperación internacional y una cultura de seguridad, y escalar los esfuerzos en la investigación de alineación.

Si bien no está claro qué tan rápido progresarán las capacidades de inteligencia artificial o qué tan rápido crecerán los riesgos catastróficos, la gravedad potencial de estas consecuencias requiere un enfoque proactivo para salvaguardar el futuro de la humanidad. A medida que nos apoyamos en el precipicio de un futuro impulsado por la IA, las elecciones que tomamos hoy podrían ser la diferencia entre cosechar los frutos de nuestra innovación o lidiar con la catástrofe.

Preguntas Frecuentes

1. ¿No deberíamos abordar los riesgos de la IA en el futuro cuando las IA realmente pueden hacer todo lo que un humano puede?

It is not necessarily the case that human-level AI is far in the future. Many top AI researchers think that human-level AI will be developed fairly soon, so urgency is warranted. Furthermore, waiting until the last second to start addressing AI risks is waiting until it's too late. Just as waiting to fully understand COVID-19 before taking any action would have been a mistake,

it is ill-advised to procrastinate on safety and wait for malicious AIs or bad actors to cause harm before taking AI risks seriously.

One might argue that since AIs cannot even drive cars or fold clothes yet, there is no need to worry. However, AIs do not need all human capabilities to pose serious threats; they only need a few specific capabilities to cause catastrophe. For example, AIs with the ability to hack computer systems or create bioweapons would pose significant risks to humanity, even if they couldn't iron a shirt. Furthermore, the development of AI capabilities has not followed an intuitive pattern where tasks that are easy for humans are the first to be mastered by AIs. Current AIs can already perform complex tasks such as writing code and designing novel drugs, even while they struggle with simple physical tasks. Like climate change and COVID-19, AI risk should be addressed proactively, focusing on prevention and preparedness rather than waiting for consequences to manifest themselves, as they may already be irreparable by that point.

2. Dado que los humanos programan IA, ¿no deberíamos ser capaces de cerrarlos si se vuelven peligrosos?

While humans are the creators of AI, maintaining control over these creations as they evolve and become more autonomous is not a guaranteed prospect. The notion that we could simply "shut them down" if they pose a threat is more complicated than it first appears.

First, consider the rapid pace at which an AI catastrophe could unfold. Analogous to preventing a rocket explosion after detecting a gas leak, or halting the spread of a virus already rampant in the population, the time between recognizing the danger and being able to prevent or mitigate it could be precariously short.

Second, over time, evolutionary forces and selection pressures could create AIs exhibiting selfish behaviors that make them more fit, such that it is harder to stop them from propagating their information. As these AIs continue to evolve and become more useful, they may become central to our societal infrastructure and daily lives, analogous to how the internet has become an essential, non-negotiable part of our lives with no

simple off-switch. They might manage critical tasks like running our energy grids, or possess vast amounts of tacit knowledge, making them difficult to replace. As we become more reliant on these AIs, we may voluntarily cede control and delegate more and more tasks to them. Eventually, we may find ourselves in a position where we lack the necessary skills or knowledge to perform these tasks ourselves. This increasing dependence could make the idea of simply "shutting them down" not just disruptive, but potentially impossible.

Similarly, some people would strongly resist or counteract attempts to shut them down, much like how we cannot permanently shut down all illegal websites or shut down Bitcoin—many people are invested in their continuation. As AIs become more vital to our lives and economies, they could develop a dedicated user base, or even a fanbase, that could actively resist attempts to restrict or shut down AIs. Likewise, consider the complications arising from malicious actors. If malicious actors have control over AIs, they could potentially use them to inflict harm. Unlike AIs under benign control, we wouldn't have an off-switch for these systems.

Next, as some AIs become more and more human-like, some may argue that these AIs should have rights. They could argue that not giving them rights is a form of slavery and is morally abhorrent. Some countries or jurisdictions may grant certain AIs rights. In fact, there is already momentum to give AIs rights. Sophia the Robot has already been granted citizenship in Saudi Arabia, and Japan granted a robot named Paro a *koseki*, or household registry, "which confirms the robot's Japanese citizenship" [135]. There may come a time when switching off an AI could be likened to murder. This would add a layer of political complexity to the notion of a simple "off-switch."

Lastly, as AIs gain more power and autonomy, they might develop a drive for "self-preservation." This would make them resistant to shutdown attempts and could allow them to anticipate and circumvent our attempts at control. Given these challenges, it's critical that we address potential AI risks proactively and put robust safeguards in place well before these problems arise.

3. ¿Por qué no podemos decirles a las IAs que sigan las Tres Leyes de Robótica de Isaac Asimov?

Asimov's laws, often highlighted in AI discussions, are insightful but inherently flawed. Indeed, Asimov himself acknowledges their limitations in his books and uses them primarily as an illustrative tool. Take the first law, for example. This law dictates that robots "may not injure a human being or, through inaction, allow a human being to come to harm," but the definition of "harm" is very nuanced. Should your home robot prevent you from leaving your house and entering traffic because it could potentially be harmful? On the other hand, if it confines you to the home, harm might befall you there as well. What about medical decisions? A given medication could have harmful side effects for some people, but not administering it could be harmful as well. Thus, there would be no way to follow this law. More importantly, the safety of AI systems cannot be ensured merely through a list of axioms or rules. Moreover, this approach would fail to address numerous technical and sociotechnical problems, including goal drift, proxy gaming, and competitive pressures. Therefore, AI safety requires a more comprehensive, proactive, and nuanced approach than simply devising a list of rules for AIs to adhere to.

4. Si las IA se vuelven más inteligentes que las personas, ¿no serían más sabias y morales? Eso significaría que no pretenderían dañarnos.

The idea of AIs becoming inherently more moral as they increase in intelligence is an intriguing concept, but rests on uncertain assumptions that can't guarantee our safety. Firstly, it assumes that moral claims can be true or false and their correctness can be discovered through reason. Secondly, it assumes that the moral claims that are really true would be beneficial for humans if AIs apply them. Thirdly, it assumes that AIs that know about morality will choose to make their decisions based on morality and not based on other considerations. An insightful parallel can be drawn to human sociopaths, who, despite their intelligence and moral awareness, do not necessarily exhibit moral inclinations or actions. This comparison illustrates that knowledge of morality

does not always lead to moral behavior. Thus, while some of the above assumptions may be true, betting the future of humanity on the claim that all of them are true would be unwise.

Assuming AIs could indeed deduce a moral code, its compatibility with human safety and wellbeing is not guaranteed. For example, AIs whose moral code is to maximize wellbeing for all life might seem good for humans at first. However, they might eventually decide that humans are costly and could be replaced with AIs that experience positive wellbeing more efficiently. AIs whose moral code is not to kill anyone would not necessarily prioritize human wellbeing or happiness, so our lives may not necessarily improve if the world begins to be increasingly shaped by and for AIs. Even AIs whose moral code is to improve the wellbeing of the worst-off in society might eventually exclude humans from the social contract, similar to how many humans view livestock. Finally, even if AIs discover a moral code that is favorable to humans, they may not act on it due to potential conflicts between moral and selfish motivations. Therefore, the moral progression of AIs is not inherently tied to human safety or prosperity.

5. ¿No se alinearían los sistemas de IA con los valores actuales perpetuarían los fracasos morales existentes?

There are plenty of moral failures in society today that we would not want powerful AI systems to perpetuate into the future. If the ancient Greeks had built powerful AI systems, they might have imbued them with many values that people today would find unethical. However, this concern should not prevent us from developing methods to control AI systems.

To achieve any value in the future, life needs to exist in the first place. Losing control over advanced AIs could constitute an existential catastrophe. Thus, uncertainty over what ethics to embed in AIs is not in tension with whether to make AIs safe.

To accommodate moral uncertainty, we should deliberately build AI systems that are adaptive and responsive to evolving moral views. As we identify moral mistakes and improve our ethical understanding, the goals we give to AIs should change accordingly—though allowing AI goals to drift unintentionally

would be a serious mistake. AIs could also help us better live by our values. For individuals, AIs could help people have more informed preferences by providing them with ideal advice [132].

Separately, in designing AI systems, we should recognize the fact of reasonable pluralism, which acknowledges that reasonable people can have genuine disagreements about moral issues due to their different experiences and beliefs [136]. Thus, AI systems should be built to respect a diverse plurality of human values, perhaps by using democratic processes and theories of moral uncertainty. Just as people today convene to deliberate on disagreements and make consensus decisions, AIs could emulate a parliament representing different stakeholders, drawing on different moral views to make real-time decisions [55, 137]. It is crucial that we deliberately design AI systems to account for safety, adaptivity, stakeholders with different values.

6. ¿No podrían los beneficios potenciales que las IA podrían traer justificar los riesgos?

The potential benefits of AI could justify the risks if the risks were negligible. However, the chance of existential risk from AI is too high for it to be prudent to rapidly develop AI. Since extinction is forever, a far more cautious approach is required. This is not like weighing the risks of a new drug against its potential side effects, as the risks are not localized but global. Rather, a more prudent approach is to develop AI slowly and carefully such that existential risks are reduced to a negligible level (e.g., under 0.001% per century).

Some influential technology leaders are accelerationists and argue for rapid AI development to barrel ahead toward a technological utopia. This techno-utopian viewpoint sees AI as the next step down a predestined path toward unlocking humanity's cosmic endowment. However, the logic of this viewpoint collapses on itself when engaged on its own terms. If one is concerned with the cosmic stakes of developing AI, we can see that even then it's prudent to bring existential risk to a negligible level. The techno-utopians suggest that delaying AI costs humanity access to a new galaxy each year, but if we go extinct, we could lose the cosmos. Thus, the prudent path is to

delay and safely prolong AI development, prioritizing risk reduction over acceleration, despite the allure of potential benefits.

7. ¿No aumentaría la atención sobre los riesgos catastróficos de las IA ahogaría los riesgos urgentes de hoy de las IA?

Focusing on catastrophic risks from AIs doesn't mean ignoring today's urgent risks; both can be addressed simultaneously, just as we can concurrently conduct research on various different diseases or prioritize mitigating risks from climate change and nuclear warfare at once. Additionally, current risks from AI are also intrinsically related to potential future catastrophic risks, so tackling both is beneficial. For example, extreme inequality can be exacerbated by AI technologies that disproportionately benefit the wealthy, while mass surveillance using AI could eventually facilitate unshakeable totalitarianism and lock-in. This demonstrates the interconnected nature of immediate concerns and long-term risks, emphasizing the importance of addressing both categories thoughtfully.

Additionally, it's crucial to address potential risks early in system development. As illustrated by Frola and Miller in their report for the Department of Defense, approximately 75 percent of the most critical decisions impacting a system's safety occur early in its development [138]. Ignoring safety considerations in the early stages often results in unsafe design choices that are highly integrated into the system, leading to higher costs or infeasibility of retrofitting safety solutions later. Hence, it is advantageous to start addressing potential risks early, regardless of their perceived urgency.

8. ¿No están trabajando muchos investigadores de IA para hacer que las IA sean seguras?

Few researchers are working to make AI safer. Currently, approximately 2 percent of papers published at top machine learning venues are safety-relevant [105]. Most of the other 98 percent focus on building more powerful AI systems more quickly. This disparity underscores the need for more balanced efforts. However, the proportion of researchers alone doesn't equate to overall safety. AI safety is a sociotechnical problem,

not just a technical problem. Thus, it requires more than just technical research. Comfort should stem from rendering catastrophic AI risks negligible, not merely from the proportion of researchers working on making AIs safe.

9. Dado que se necesitan miles de años para producir cambios significativos, ¿por qué tenemos que preocuparnos de que la evolución sea una fuerza impulsora en el desarrollo de la IA?

Although the biological evolution of humans is slow, the evolution of other organisms, such as fruit flies or bacteria, can be extremely quick, demonstrating the diverse time scales at which evolution operates. The same rapid evolutionary changes can be observed in non-biological structures like software, which evolve much faster than biological entities. Likewise, one could expect AIs to evolve very quickly as well. The rate of AI evolution may be propelled by intense competition, high variation due to diverse forms of AIs and goals given to them, and the ability of AIs to rapidly adapt. Consequently, intense evolutionary pressures may be a driving force in the development of AIs.

10. ¿No necesitarían las IA tener un impulso de búsqueda de energía para representar un riesgo grave?

While power-seeking AI poses a risk, it is not the only scenario that could potentially lead to catastrophe. Malicious or reckless use of AIs can be equally damaging without the AI itself seeking power. Additionally, AIs might engage in harmful actions through proxy gaming or goal drift without intentionally seeking power. Furthermore, society's trend toward automation, driven by competitive pressures, is gradually increasing the influence of AIs over humans. Hence, the risk does not solely stem from AIs seizing power, but also from humans ceding power to AIs.

11. ¿No es la combinación de la inteligencia humana y la IA superior a la IA sola, por lo que no hay necesidad de preocuparse por el desempleo o los humanos que se vuelven irrelevantes?

Si bien es cierto que los equipos de computación humana han superado a las computadoras solas en el pasado, estos han sido

fenómenos temporales. Por ejemplo, el "ajedrez cyborg" es una forma de ajedrez donde los humanos y las computadoras trabajan juntos, que históricamente era superior a los humanos o las computadoras solas. Sin embargo, los avances en los algoritmos de ajedrez por ordenador han erosionado la ventaja de los equipos humano-ordenador hasta tal punto que podría decirse que ya no hay ninguna ventaja en comparación con los ordenadores solos. Para tomar un ejemplo más simple, nadie enfrentaría a un humano contra una simple calculadora para una división larga. Una progresión similar puede ocurrir con las IA. Puede haber una fase provisional en la que los humanos y las IA puedan trabajar juntos de manera efectiva, pero la tendencia sugiere que las IA por sí solas podrían eventualmente superar a los humanos en varias tareas sin beneficiarse más de la asistencia humana.

12. El desarrollo de la IA parece imparable. ¿No lo ralentizaría dramáticamente o detenerlo requeriría algo así como un régimen de vigilancia global invasiva?

AI development primarily relies on high-end chips called GPUs, which can be feasibly monitored and tracked, much like uranium. Additionally, the computational and financial investments required to develop frontier AIs are growing exponentially, resulting in a small number of actors who are capable of acquiring enough GPUs to develop them. Therefore, managing AI growth doesn’t necessarily require invasive global surveillance, but rather a systematic tracking of high-end GPU usage.

Thank you!

CAIS es una organización sin fines de lucro de seguridad de IA. Nuestra misión es reducir los riesgos a escala social de la inteligencia artificial.