Descripción general de los riesgos catastróficos de la IA

Existe una variedad de riesgos derivados de los sistemas de IA, incluidos tanto los riesgos actuales como los que pueden surgir en el futuro cercano. Estos incluyen riesgos de resultados catastróficos, que pueden dividirse en riesgos de uso malicioso, carreras de IA, riesgos organizacionales e IA no autorizadas.

1.1 Introducción

En este capítulo, ofrecemos una descripción breve e informal de muchos de los principales riesgos a escala social derivados de la IA, centrándonos en los riesgos de la IA que podrían conducir a resultados sociales muy graves o incluso catastróficos. Esto proporciona algunos antecedentes y motivación antes de discutir desafíos específicos con más profundidad y rigor en los siguientes capítulos.

El mundo tal como lo conocemos no es normal. Damos por sentado que podemos hablar instantáneamente con personas que se encuentran a miles de kilómetros de distancia, volar al otro lado del mundo en menos de un día y acceder a vastas montañas de conocimiento acumulado en dispositivos que llevamos en el bolsillo. Estas realidades parecían descabelladas hace décadas y habrían sido inconcebibles para las personas que vivieron hace siglos. Las formas en que vivimos, trabajamos, viajamos y nos comunicamos sólo han sido posibles durante una pequeña fracción de la historia de la humanidad.

Sin embargo, cuando miramos el panorama más amplio, emerge un patrón más amplio: la aceleración del desarrollo. Pasaron cientos de miles de años entre la aparición del Homo sapiens en la Tierra y la revolución agrícola. Luego, pasaron miles de años antes de la revolución industrial. Ahora, apenas siglos después, está comenzando la revolución de la inteligencia artificial (IA). La marcha de la historia no es constante: se está acelerando rápidamente.

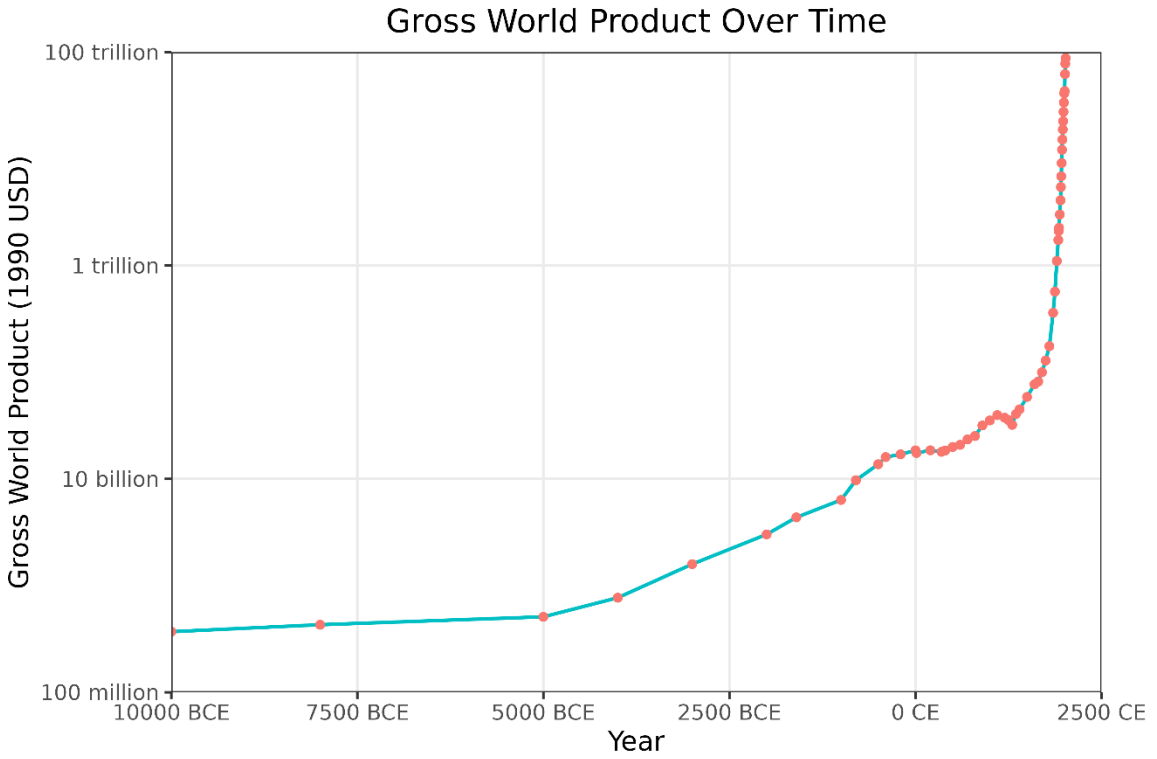

Figura 1.1 La producción mundial ha crecido rápidamente a lo largo de la historia de la humanidad. La IA podría promover esta tendencia, catapultando a la humanidad a un nuevo período de cambios sin precedentes.

Podemos capturar esta tendencia cuantitativamente en la Figura 1.1, que muestra cómo el producto mundial bruto estimado ha cambiado a lo largo del tiempo [1], [2]. El crecimiento hiperbólico que describe podría explicarse por el hecho de que, a medida que avanza la tecnología, la tasa de avance tecnológico también tiende a aumentar. Dotadas de nuevas tecnologías, las personas pueden innovar más rápido que antes. Por lo tanto, la brecha de tiempo entre cada acontecimiento histórico se reduce.

Es el rápido ritmo de desarrollo, así como la sofisticación de nuestra tecnología, lo que hace que el presente sea un momento sin precedentes en la historia de la humanidad. Hemos llegado a un punto en el que los avances tecnológicos pueden transformar el mundo más allá del reconocimiento en el transcurso de una vida humana. Por ejemplo, las personas que han vivido la creación de Internet pueden recordar una época en la que nuestro mundo ahora conectado digitalmente habría parecido ciencia ficción.

Desde una perspectiva histórica, parece posible que la misma cantidad de desarrollo pueda ahora condensarse en un período de tiempo aún más corto. Puede que no estemos seguros de que esto vaya a ocurrir, pero tampoco podemos descartarlo. Por lo tanto, nos preguntamos: ¿qué nueva tecnología podría marcar el comienzo de la próxima gran aceleración? A la luz de los avances recientes, la IA parece un candidato cada vez más plausible. Quizás, a medida que la IA siga volviéndose más poderosa, podría conducir a un cambio cualitativo en el mundo, más profundo que cualquiera de

los que hayamos experimentado hasta ahora. Podría ser el período más impactante de la historia, aunque también podría ser el último.

Aunque los avances tecnológicos a menudo han mejorado la vida de las personas, debemos recordar que, a medida que nuestra tecnología crece en poder, también lo hace su potencial destructivo. Consideremos la invención de las armas nucleares. El siglo pasado, por primera vez en la historia de nuestra especie, la humanidad poseyó la capacidad de destruirse a sí misma y el mundo de repente se volvió mucho más frágil.

Nuestra nueva vulnerabilidad se reveló con una claridad desconcertante durante la Guerra Fría. Un sábado de octubre de 1962, la crisis de los misiles cubanos se estaba saliendo de control. Los buques de guerra estadounidenses que imponían el bloqueo a Cuba detectaron un submarino soviético e intentaron obligarlo a salir a la superficie lanzando cargas de profundidad de bajo explosivo. El submarino estaba fuera de contacto por radio y su tripulación no tenía idea de si la Tercera Guerra Mundial ya había comenzado. Un ventilador roto elevó la temperatura hasta 140 ◦F en algunas partes del submarino, lo que provocó que los miembros de la tripulación quedaran inconscientes cuando las cargas de profundidad explotaron cerca.

El submarino llevaba un torpedo con armas nucleares, cuyo lanzamiento requería el consentimiento tanto del capitán como del oficial político. Ambos lo proporcionaron. En cualquier otro submarino en aguas cubanas ese día, ese torpedo se habría lanzado, y es posible que hubiera seguido una tercera guerra mundial nuclear. Afortunadamente, en el submarino también iba un hombre llamado Vasili Arkhipov. Arkhipov era el comandante de toda la flotilla y por pura suerte se encontraba en ese submarino en particular. Convenció al capitán de su ira y lo convenció de que esperara nuevas órdenes de Moscú. Evitó una guerra nuclear y salvó millones o miles de millones de vidas, y posiblemente la civilización misma.

Carl Sagan observó una vez: “Si continuamos acumulando sólo poder y no sabiduría, seguramente nos destruiremos a nosotros mismos” [3]. Sagan tenía razón: no estábamos preparados para el poder de las armas nucleares. En general, ha sido suerte más que sabiduría lo que ha salvado a la humanidad de la aniquilación nuclear, con múltiples casos registrados de un solo individuo que evitó una guerra nuclear a gran escala.

La IA está ahora a punto de convertirse en una poderosa tecnología con un potencial destructivo similar al de las armas nucleares. No queremos que se repita la crisis de los misiles cubanos. No queremos deslizarnos hacia un momento de peligro en el que nuestra supervivencia dependa de la suerte en lugar de la capacidad de utilizar esta tecnología sabiamente. En cambio, debemos trabajar de manera proactiva para mitigar los riesgos que plantea. Esto requiere una mejor comprensión de lo que podría salir mal y qué hacer al respecto.

Afortunadamente, los sistemas de inteligencia artificial aún no están lo suficientemente avanzados como para contribuir a todos los riesgos que analizamos. Pero eso es poco consuelo en una época en la que el desarrollo de la IA avanza a un ritmo impredecible y sin precedentes. Consideramos los riesgos que surgen tanto de las IA actuales como de las que probablemente existan en el futuro cercano. Es

posible que si esperamos a que se desarrollen sistemas más avanzados antes de tomar medidas, sea demasiado tarde.

En este capítulo, exploraremos varias formas en que las IA poderosas podrían provocar eventos catastróficos con consecuencias devastadoras para un gran número de personas. También discutiremos cómo las IA podrían presentar riesgos existenciales: catástrofes de las que la humanidad no podría recuperarse. El riesgo más obvio es la extinción, pero hay otros resultados, como la creación de una sociedad distópica permanente, que también constituiría una catástrofe existencial. Describimos muchas catástrofes posibles, algunas de las cuales son más probables que otras y algunas son mutuamente incompatibles entre sí. Este enfoque está motivado por los principios de la gestión de riesgos. Damos prioridad a preguntar "¿qué podría salir mal?" en lugar de esperar reactivamente a que se produzcan catástrofes. Esta mentalidad proactiva nos permite anticipar y mitigar riesgos catastróficos antes de que sea demasiado tarde.

Para ayudar a orientar la discusión, descomponemos los riesgos catastróficos de las IA en cuatro fuentes de riesgo que justifican la intervención:

- **Uso malicioso** : actores maliciosos que utilizan IA para causar devastación a gran escala.
- **Carrera de IA** : presiones competitivas que podrían llevarnos a implementar IA de maneras inseguras, a pesar de que esto no es lo mejor para nadie.
- **Riesgos organizacionales** : Accidentes derivados de la complejidad de las IA y las organizaciones que las desarrollan.
- **IA rebeldes** : el problema de controlar una tecnología más inteligente que nosotros.

Estas cuatro secciones Uso malicioso, Carreras de IA, Riesgos organizacionales e IA no autorizada describen causas de riesgos de IA que son intencionales, ambientales/estructurales, accidentales e internos, respectivamente [4]. Los riesgos que se describen brevemente en este capítulo se analizan con mayor profundidad en el resto de este libro.

En este capítulo describiremos cómo ejemplos concretos y a pequeña escala de cada riesgo podrían derivar en resultados catastróficos. También incluimos historias hipotéticas para ayudar a los lectores a conceptualizar los diversos procesos y dinámicas discutidos en cada sección. Esperamos que esta encuesta sirva como una introducción práctica para los lectores interesados en conocer y mitigar los riesgos catastróficos de la IA.

Referencias

[1] D. M. Roodman, “On the probability distribution of long-term changes in the growth rate of the global economy: An outside view.” 2020.
[2] T. Davidson, “Could advanced AI drive explosive economic growth?” 2021.
[3] C. Sagan, *Pale blue dot: A vision of the human future in space*. New York: Random House, 1994.
[4] R. V. Yampolskiy, “Taxonomy of pathways to dangerous artificial intelligence,” in *AAAI workshop: AI, ethics, and society*, 2016.

1.2 Uso malintencionado

Los sistemas de inteligencia artificial podrían utilizarse con fines maliciosos como el terrorismo, la manipulación y la desinformación, o para afianzar un estado totalitario.

La mañana del 20 de marzo de 1995, cinco hombres entraron en el metro de Tokio. Después de abordar líneas de metro separadas, continuaron durante varias paradas antes de dejar las bolsas que llevaban y salir. Un líquido inodoro e incoloro dentro de las bolsas comenzó a vaporizarse. En cuestión de minutos, los viajeros comenzaron a asfixiarse y vomitar. Los trenes continuaron hacia el corazón de Tokio, y los pasajeros enfermos abandonaron los vagones en cada estación. Los vapores se esparcieron en cada parada, ya sea emanando de los autos contaminados o a través del contacto con la ropa y los zapatos de las personas. Al final del día, 13 personas yacían muertas y 5.800 gravemente heridas. El grupo responsable del ataque fue el culto religioso Aum Shinrikyo [1]. ¿Su motivo para asesinar a personas inocentes? Provocar el fin del mundo.

Las nuevas y poderosas tecnologías ofrecen enormes beneficios potenciales, pero también conllevan el riesgo de empoderar a actores maliciosos para causar daños generalizados. Siempre habrá quienes tengan las peores intenciones, y las IA podrían proporcionarles una herramienta formidable para lograr sus objetivos. Además, a medida que avanza la tecnología de IA, el uso malicioso grave podría desestabilizar a la sociedad, aumentando la probabilidad de que se produzcan otros riesgos.

En esta sección, exploraremos las diversas formas en que el uso malicioso de IA avanzadas podría plantear riesgos catastróficos. Estas incluyen diseñar armas bioquímicas, desatar IA rebeldes, utilizar IA persuasivas para difundir propaganda y erosionar la realidad consensuada, y aprovechar la censura y la vigilancia masiva para concentrar el poder de manera irreversible. Concluiremos discutiendo posibles estrategias para mitigar los riesgos asociados con el uso malicioso de la IA.

Los actores unilaterales aumentan considerablemente los riesgos de uso malicioso. En los casos en que numerosos actores tienen acceso a una tecnología poderosa o a información peligrosa que podría usarse con fines dañinos, solo se necesita un individuo para causar una devastación significativa. Los propios actores maliciosos son el ejemplo más claro de esto, pero la imprudencia puede ser igualmente peligrosa. Por ejemplo, un solo equipo de investigación podría estar entusiasmado con la idea de abrir un sistema de IA con capacidades de investigación biológica, lo que aceleraría la investigación y potencialmente salvaría vidas, pero esto también podría aumentar el riesgo de uso malicioso si el sistema de IA pudiera reutilizarse para desarrollar armas biológicas. En situaciones como ésta, el resultado puede ser determinado por el grupo de investigación con menor aversión al riesgo. Si sólo un grupo de investigación piensa que los beneficios superan los riesgos, podría actuar unilateralmente y decidir el resultado incluso si la mayoría de los demás no están de acuerdo. Y si se equivocan y alguien decide desarrollar un arma biológica, sería demasiado tarde para revertir el rumbo.

Por defecto, las IA avanzadas pueden aumentar la capacidad destructiva tanto de los más poderosos como de la población en general. Por lo tanto, el creciente potencial de las IA para empoderar a actores maliciosos es una de las amenazas más graves a las que se enfrentará la humanidad en las próximas décadas. Los ejemplos que damos en esta sección son sólo los que podemos prever. Es posible que las IA puedan ayudar en la creación de nuevas tecnologías peligrosas que actualmente no podemos imaginar, lo que aumentaría aún más los riesgos de uso malicioso.

1.2.1 Bioterrorismo

El rápido avance de la tecnología de inteligencia artificial aumenta el riesgo de bioterrorismo. Las IA con conocimientos de bioingeniería podrían facilitar la creación de nuevas armas biológicas y reducir las barreras para la obtención de dichos agentes. Las pandemias diseñadas a partir de armas biológicas asistidas por IA plantean un desafío único, ya que los atacantes tienen una ventaja sobre los defensores y podrían constituir una amenaza existencial para la humanidad. Ahora examinaremos estos riesgos y cómo las IA podrían exacerbar los desafíos en la gestión del bioterrorismo y las pandemias diseñadas.

Las pandemias creadas por bioingeniería presentan una nueva amenaza. Los agentes biológicos, incluidos virus y bacterias, han causado algunas de las catástrofes más devastadoras de la historia. Se cree que la Peste Negra mató a más humanos que cualquier otro evento en la historia, unos asombrosos y terribles 200 millones, el equivalente a cuatro mil millones de muertes en la actualidad. Si bien los avances contemporáneos en ciencia y medicina han logrado grandes avances en la mitigación de los riesgos asociados con las pandemias naturales, las pandemias diseñadas podrían diseñarse para ser más letales o fácilmente transmisibles que las pandemias naturales, presentando una nueva amenaza que podría igualar o incluso superar la devastación provocada por los cambios históricos. plagas más mortíferas [2].

La humanidad tiene una larga y oscura historia de convertir patógenos en armas, con registros que se remontan al año 1320 a. C. que describen una guerra en Asia Menor en la que ovejas infectadas fueron conducidas a través de la frontera para propagar la tularemia [3]. Durante el siglo XX, se sabe que 15 países desarrollaron programas de armas biológicas, entre ellos Estados Unidos, la URSS, el Reino Unido y Francia. Al igual que las armas químicas, las armas biológicas se han convertido en un tabú entre la comunidad internacional. Si bien algunos actores estatales continúan operando programas de armas biológicas [4], un riesgo más significativo puede provenir de actores no estatales como Aum Shinrikyo, ISIS o simplemente individuos perturbados. Debido a los avances en la inteligencia artificial y la biotecnología, las herramientas y el conocimiento necesarios para diseñar patógenos con capacidades mucho más allá de los programas de armas biológicas de la era de la Guerra Fría se democratizarán rápidamente.

La biotecnología avanza rápidamente y se vuelve más accesible. Hace unas décadas, la capacidad de sintetizar nuevos virus estaba limitada a un puñado de los mejores científicos que trabajaban en laboratorios avanzados. Hoy se estima que hay 30.000 personas con el talento, la formación y el acceso a la tecnología para crear nuevos patógenos [2]. Esta cifra podría expandirse rápidamente. La síntesis de

genes, que permite la creación de agentes biológicos personalizados, ha bajado vertiginosamente de precio, y su costo se reduce a la mitad aproximadamente cada 15 meses [5]. Además, con la llegada de las máquinas de síntesis de ADN de mesa de trabajo, el acceso será mucho más fácil y podría evitar los esfuerzos de detección de síntesis de genes existentes, lo que complica el control de la difusión de dicha tecnología [6]. Las posibilidades de que una pandemia creada por bioingeniería mate a millones, tal vez miles de millones, son proporcionales al número de personas con las habilidades y el acceso a la tecnología para sintetizarlas. Con los asistentes de IA, muchas más personas podrían tener las habilidades necesarias, aumentando así los riesgos en órdenes de magnitud.

Las IA podrían utilizarse para acelerar el descubrimiento de armas químicas y biológicas nuevas y más mortíferas. En 2022, los investigadores tomaron un sistema de inteligencia artificial diseñado para crear nuevos medicamentos mediante la generación de moléculas terapéuticas no tóxicas y lo modificaron para recompensar, en lugar de penalizar, la toxicidad [7]. Después de este simple cambio, en seis horas, generó por sí solo 40.000 candidatos a agentes de guerra química. No sólo diseñó sustancias químicas mortales conocidas, incluido el VX, sino también moléculas novedosas que pueden ser más letales que cualquier agente de guerra química descubierto hasta ahora. En el campo de la biología, las IA ya han superado las capacidades humanas en la predicción de la estructura de las proteínas [8] y han contribuido a la síntesis de esas proteínas [9]. Se podrían utilizar métodos similares para crear armas biológicas y desarrollar patógenos que sean más mortales, más transmisibles y más difíciles de tratar que cualquier cosa vista antes.

Las IA agravan la amenaza de las pandemias creadas por bioingeniería. Las IA aumentarán el número de personas que podrían cometer actos de bioterrorismo. Las IA de uso general como ChatGPT son capaces de sintetizar conocimiento experto sobre los patógenos más mortíferos conocidos, como la influenza y la viruela, y proporcionar instrucciones paso a paso sobre cómo una persona podría crearlos mientras evade los protocolos de seguridad [10]. Las versiones futuras de las IA podrían ser aún más útiles para los bioterroristas potenciales cuando las IA sean capaces de sintetizar información en técnicas, procesos y conocimientos que no están disponibles explícitamente en ningún lugar de Internet. Las autoridades de salud pública pueden responder a estas amenazas con medidas de seguridad, pero en el bioterrorismo, el atacante tiene la ventaja. La naturaleza exponencial de las amenazas biológicas significa que un solo ataque podría extenderse al mundo entero antes de que se pueda montar una defensa eficaz. Sólo 100 días después de ser detectada y secuenciada, la variante omicrón del COVID-19 había infectado una cuarta parte de Estados Unidos y la mitad de Europa [2]. Las cuarentenas y los confinamientos instituidos para reprimir la pandemia de COVID-19 provocaron una recesión mundial y aún así no pudieron evitar que la enfermedad matara a millones de personas en todo el mundo.

En resumen, las IA avanzadas podrían constituir un arma de destrucción masiva en manos de terroristas, al facilitarles el diseño, la síntesis y la propagación de nuevos patógenos mortales. Al reducir la experiencia técnica requerida y aumentar la letalidad y transmisibilidad de los patógenos, las IA podrían permitir que actores maliciosos causen una catástrofe global al desencadenar pandemias.

1.2.2 Liberación de agentes de IA

Muchas tecnologías son *herramientas* que los humanos utilizamos para alcanzar nuestros objetivos, como martillos, tostadoras y cepillos de dientes. Pero las IA se construyen cada vez más como *agentes* que toman acciones de forma autónoma en el mundo para perseguir objetivos abiertos. A los agentes de IA se les pueden asignar objetivos como ganar juegos, obtener ganancias en el mercado de valores o conducir un automóvil a un destino. Por lo tanto, los agentes de IA plantean un riesgo único: las personas podrían construir IA que persigan objetivos peligrosos.

Los actores malintencionados podrían crear intencionalmente IA no autorizadas. Un mes después del lanzamiento de GPT-4, un proyecto de código abierto pasó por alto los filtros de seguridad de la IA y la convirtió en un agente autónomo de IA con instrucciones de "destruir a la humanidad", "establecer el dominio global" y "alcanzar la inmortalidad". Apodado ChaosGPT , la IA recopiló investigaciones sobre armas nucleares y envió tweets tratando de influir en otros. Afortunadamente, ChaosGPT fue simplemente una advertencia dado que carecía de la capacidad de formular con éxito planes a largo plazo, piratear computadoras y sobrevivir y propagarse. Sin embargo, dado el rápido ritmo del desarrollo de la IA, ChaosGPT ofreció una idea de los riesgos que podrían plantear las IA no autorizadas más avanzadas en el futuro cercano.

Es posible que muchos grupos quieran liberar IA o hacer que las IA desplacen a la humanidad. Simplemente desatar IA rebeldes, como una versión más sofisticada de ChaosGPT, podría lograr una destrucción masiva, incluso si a esas IA no se les ordena explícitamente que dañen a la humanidad. Hay una variedad de creencias que pueden impulsar a individuos y grupos a hacerlo. Una ideología que podría representar una amenaza única en este sentido es el “aceleracionismo”. Esta ideología busca acelerar el desarrollo de la IA lo más rápido posible y se opone a las restricciones al desarrollo o la proliferación de las IA. Este sentimiento es común entre muchos investigadores destacados de la IA y líderes tecnológicos, algunos de los cuales están compitiendo intencionalmente para construir IA más inteligentes que los humanos. Según Larry Page, cofundador de Google, las IA son las herederas legítimas de la humanidad y el siguiente paso de la evolución cósmica. También ha expresado el sentimiento de que el hecho de que los seres humanos mantengan el control sobre las IA es “ especista ” [11]. Jürgen Schmidhuber , un eminente científico de IA, argumentó que “A largo plazo, los humanos no seguirán siendo la corona de la creación... Pero eso está bien porque todavía hay belleza, grandeza y grandeza en darte cuenta de que eres una pequeña parte de un esquema mucho más amplio que está llevando al universo desde una menor complejidad hacia una mayor complejidad” [12]. Richard Sutton, otro destacado científico de la IA, al hablar de una IA más inteligente que la humana preguntó "¿por qué los más inteligentes no deberían volverse poderosos?" y piensa que el desarrollo de la superinteligencia será un logro “más allá de la humanidad, más allá de la vida, más allá del bien y del mal” [13]. Sostiene que “la sucesión en la IA es inevitable” y, si bien “podrían desplazarnos de la existencia”, “no debemos resistirnos a la sucesión” [14].

Hay varios grupos importantes que podrían querer utilizar IA para causar daño intencionalmente. Por ejemplo, los sociópatas y psicópatas constituyen alrededor del 3 por ciento de la población [15]. En el futuro, es posible que las personas cuyos

medios de vida sean destruidos por la automatización de la IA se sientan resentidas y algunas quieran tomar represalias. Hay muchos casos en los que personas aparentemente mentalmente estables sin antecedentes de locura o violencia de repente se lanzan a disparar o colocan una bomba con la intención de dañar a tantas personas inocentes como sea posible. También podemos esperar que personas bien intencionadas hagan la situación aún más desafiante. A medida que las IA avanzan, podrían ser compañeras ideales: saber cómo brindar comodidad, ofrecer consejos cuando sea necesario y nunca exigir nada a cambio. Inevitablemente, las personas desarrollarán vínculos emocionales con los chatbots y algunos exigirán que se les concedan derechos o se vuelvan autónomos.

En resumen, liberar IA poderosas y permitirles tomar acciones independientemente de los humanos podría conducir a una catástrofe. Hay muchas razones por las que la gente podría perseguir esto, ya sea por el deseo de causar daño, una creencia ideológica en la aceleración tecnológica o la convicción de que las IA deberían tener los mismos derechos y libertades que los humanos.

1.2.3 IA persuasivas

La propagación deliberada de desinformación ya es un problema grave, que reduce nuestra comprensión compartida de la realidad y polariza opiniones. Las IA podrían utilizarse para exacerbar gravemente este problema generando desinformación personalizada a mayor escala que antes. Además, a medida que las IA mejoren a la hora de predecir y estimular nuestro comportamiento, serán más capaces de manipularnos. Ahora discutiremos cómo actores maliciosos podrían aprovechar las IA para crear una sociedad fracturada y disfuncional.

Las IA podrían contaminar el ecosistema de la información con mentiras motivadas. A veces las ideas se difunden no porque sean ciertas, sino porque sirven a los intereses de un grupo en particular. El “periodismo amarillo” fue acuñado como una referencia peyorativa a los periódicos que defendían la guerra entre España y Estados Unidos a finales del siglo XIX, porque creían que las historias sensacionalistas de guerra aumentarían sus ventas [16]. Cuando las fuentes de información pública están inundadas de falsedades, las personas a veces caen presa de mentiras o llegan a desconfiar de las narrativas dominantes, lo cual socava la integridad social.

Desafortunadamente, las IA podrían agravar drásticamente estos problemas existentes. En primer lugar, las IA podrían utilizarse para generar desinformación única y personalizada a gran escala. Si bien ya existen muchos robots de redes sociales [17], algunos de los cuales existen para difundir desinformación, históricamente han sido administrados por humanos o generadores de texto primitivos. Los últimos sistemas de inteligencia artificial no necesitan que los humanos generen mensajes personalizados, nunca se cansan y podrían interactuar con millones de usuarios a la vez [18].

Las IA pueden explotar la confianza de los usuarios. Cientos de miles de personas ya pagan por chatbots comercializados como amantes y amigos [19], y el suicidio de un hombre se ha atribuido en parte a interacciones con un chatbot [20]. A medida que las IA se parecen cada vez más a los humanos, las personas establecerán cada vez

más relaciones con ellas y llegarán a confiar en ellas. Las IA que recopilan información personal mediante la construcción de relaciones o accediendo a una gran cantidad de datos personales, como la cuenta de correo electrónico o los archivos personales de un usuario, podrían aprovechar esa información para mejorar la persuasión. Los actores poderosos que controlan esos sistemas podrían explotar la confianza de los usuarios entregando desinformación personalizada directamente a través de los "amigos" de las personas.

Las IA podrían centralizar el control de la información confiable. Aparte de democratizar la desinformación, las IA podrían centralizar la creación y difusión de información confiable. Solo unos pocos actores tienen las habilidades técnicas y los recursos para desarrollar sistemas de IA de vanguardia, y podrían utilizar estas IA para difundir sus narrativas preferidas. Alternativamente, si las IA son ampliamente accesibles, esto podría dar lugar a una desinformación generalizada, y la gente confiaría sólo en un pequeño puñado de fuentes autorizadas [21]. En ambos escenarios, habría menos fuentes de información confiable y una pequeña porción de la sociedad controlaría las narrativas populares.

La censura de la IA podría centralizar aún más el control de la información. Esto podría comenzar con buenas intenciones, como el uso de IA para mejorar la verificación de hechos y ayudar a las personas a evitar ser víctimas de narrativas falsas. Esto no necesariamente resolvería el problema, ya que la desinformación persiste hoy a pesar de la presencia de verificadores de datos.

Peor aún, las supuestas "IA de verificación de hechos" podrían ser diseñadas por gobiernos autoritarios y otros para suprimir la difusión de información verdadera. Estas IA podrían diseñarse para corregir los conceptos erróneos más comunes, pero proporcionar información incorrecta sobre algunos temas delicados, como las violaciones de derechos humanos cometidas por ciertos países. Pero incluso si las IA de verificación de datos funcionan según lo previsto, el público podría eventualmente volverse completamente dependiente de ellas para determinar la verdad, reduciendo la autonomía de las personas y haciéndolas vulnerables a fallas o ataques a esos sistemas.

En un mundo con sistemas de IA persuasivos generalizados, las creencias de las personas podrían estar determinadas casi por completo por con qué sistemas de IA interactúan más. Sin saber en quién confiar, la gente podía retirarse aún más a enclaves ideológicos, temiendo que cualquier información procedente de fuera de esos enclaves pudiera ser una mentira sofisticada. Esto erosionaría la realidad del consenso, la capacidad de las personas para cooperar con otras, participar en la sociedad civil y abordar problemas de acción colectiva. Esto también reduciría nuestra capacidad de conversar como especie sobre cómo mitigar los riesgos existenciales de las IA.

En resumen, las IA podrían crear desinformación personalizada y altamente efectiva a una escala sin precedentes, y podrían ser particularmente persuasivas para las personas con las que han establecido relaciones personales. En manos de muchas personas, esto podría crear una avalancha de desinformación que debilita a la sociedad humana, pero, si se mantiene en manos de unos pocos, podría permitir a los gobiernos controlar las narrativas para sus propios fines.

1.2.4 Concentración de poder

Hemos discutido varias formas en que individuos y grupos podrían usar IA para causar daños generalizados, a través del bioterrorismo; liberar IA poderosas e incontroladas; y desinformación. Para mitigar estos riesgos, los gobiernos podrían implementar una vigilancia intensa y tratar de mantener las IA en manos de una minoría confiable. Sin embargo, esta reacción podría fácilmente convertirse en una sobrecorrección, allanando el camino para un régimen totalitario arraigado que quedaría atrapado por el poder y la capacidad de las IA. Este escenario representa una forma de uso indebido “de arriba hacia abajo”, a diferencia del uso indebido “de abajo hacia arriba” por parte de los ciudadanos, y en casos extremos podría culminar en una civilización distópica arraigada.

Las IA podrían conducir a una concentración de poder extrema y quizás irreversible. Las capacidades persuasivas de las IA, combinadas con su potencial para la vigilancia y el avance de armas autónomas, podrían permitir que pequeños grupos de actores “fijen” su control sobre la sociedad, tal vez de forma permanente. Para funcionar eficazmente, las IA requieren un amplio conjunto de componentes de infraestructura, que no están distribuidos equitativamente, como centros de datos , potencia informática y big data. Quienes controlan sistemas poderosos pueden utilizarlos para reprimir la disidencia, difundir propaganda y desinformación y promover de otro modo sus objetivos, lo que puede ser contrario al bienestar público.

Las IA pueden afianzar un régimen totalitario. En manos del Estado, las IA pueden provocar la erosión de las libertades civiles y los valores democráticos en general. Las IA podrían permitir a los gobiernos totalitarios recopilar, procesar y actuar de manera eficiente sobre un volumen de información sin precedentes, permitiendo a un grupo cada vez más pequeño de personas vigilar y ejercer un control total sobre la población sin la necesidad de reclutar a millones de ciudadanos para que sirvan como funcionarios gubernamentales dispuestos. En general, a medida que el poder y el control se alejan del público y se acercan a las elites y los líderes, los gobiernos democráticos son muy vulnerables a un retroceso totalitario. Además, las IA podrían hacer que los regímenes totalitarios sean mucho más duraderos; una de las principales formas en que se han derrocado regímenes de este tipo anteriormente es en momentos de vulnerabilidad como la muerte de un dictador, pero las IA, que serían difíciles de “matar”, podrían brindar mucha más continuidad al liderazgo, brindando pocas oportunidades de reforma.

Las IA pueden afianzar el poder corporativo a expensas del bien público. Las corporaciones han presionado durante mucho tiempo para debilitar las leyes y políticas que restringen sus acciones y su poder, todo ello al servicio de las ganancias. Las corporaciones que controlan potentes sistemas de inteligencia artificial pueden utilizarlos para manipular a los clientes para que gasten más en sus productos, incluso en detrimento de su propio bienestar. La concentración de poder e influencia que podrían permitir las IA podría permitir a las corporaciones ejercer un control sin precedentes sobre el sistema político y ahogar por completo las voces de los ciudadanos. Esto podría ocurrir incluso si los creadores de estos sistemas saben que sus sistemas son interesados o perjudiciales para otros, ya que tendrían incentivos para reforzar su poder y evitar distribuir el control.

Además del poder, fijar ciertos valores puede restringir el progreso moral de la humanidad. Es peligroso permitir que cualquier conjunto de valores quede permanentemente arraigado en la sociedad. Por ejemplo, los sistemas de inteligencia artificial han aprendido puntos de vista racistas y sexistas [22], y una vez que se aprenden esos puntos de vista, puede resultar difícil eliminarlos por completo. Además de los problemas que sabemos que existen en nuestra sociedad, puede haber algunos que todavía no conocemos. Así como aborrecemos algunas opiniones morales ampliamente sostenidas en el pasado, la gente en el futuro puede querer dejar atrás las opiniones morales que tenemos hoy, incluso aquellas con las que actualmente no vemos ningún problema. Por ejemplo, los defectos morales en los sistemas de IA serían aún peores si los sistemas de IA hubieran sido entrenados en la década de 1960, y muchas personas en ese momento no habrían visto ningún problema en eso. Es posible que hoy, sin saberlo, estemos perpetuando catástrofes morales [23]. Por lo tanto, cuando las IA avanzadas emergen y transforman el mundo, existe el riesgo de que sus objetivos se fijen o perpetúen defectos en los valores actuales. Si las IA no están diseñadas para aprender y actualizar continuamente su comprensión de los valores sociales, pueden perpetuar o reforzar los defectos existentes en sus procesos de toma de decisiones en el futuro.

En resumen, si bien mantener poderosas IA en manos de unos pocos podría reducir los riesgos del terrorismo, podría exacerbar aún más la desigualdad de poder si los gobiernos y las corporaciones hacen un mal uso de ellas. Esto podría conducir a un gobierno totalitario y a una intensa manipulación del público por parte de las corporaciones, y podría fijar los valores actuales, impidiendo cualquier mayor progreso moral.

Historia: Bioterrorismo

La siguiente es una historia hipotética ilustrativa para ayudar a los lectores a imaginar algunos de estos riesgos. Sin embargo, esta historia es algo vaga para reducir el riesgo de inspirar acciones maliciosas basadas en ella.

Una startup de biotecnología está causando sensación en la industria con su modelo de bioingeniería impulsado por IA. La compañía ha hecho afirmaciones audaces de que esta nueva tecnología revolucionará la medicina gracias a su capacidad de crear curas para enfermedades conocidas y desconocidas. Sin embargo, la empresa generó cierta controversia cuando decidió entregar el programa a investigadores aprobados de la comunidad científica. Sólo unas semanas después de su decisión de hacer que el modelo fuera de código abierto de forma limitada, el modelo completo se filtró en Internet para que todos lo vieran. Sus críticos señalaron que el modelo podría reutilizarse para diseñar patógenos letales y afirmaron que la filtración proporcionó a los malos actores una herramienta poderosa para causar una destrucción generalizada, lo que lo abrió a abusos sin salvaguardias establecidas.

Desconocido para el público, un grupo extremista ha estado trabajando durante años para diseñar un nuevo virus diseñado para matar a un gran

número de personas. Sin embargo, dada su falta de experiencia, estos esfuerzos hasta ahora no han tenido éxito. Cuando se filtra el nuevo sistema de IA, el grupo lo reconoce inmediatamente como una herramienta potencial para diseñar el virus y sortear obstáculos legales y de seguimiento para obtener las materias primas necesarias. El sistema de inteligencia artificial diseña con éxito exactamente el tipo de virus que esperaba el grupo extremista. También proporciona instrucciones paso a paso sobre cómo sintetizar grandes cantidades del virus y sortear cualquier obstáculo para su propagación. Con el virus sintetizado en la mano, el grupo extremista diseña un plan para liberar el virus en varios lugares cuidadosamente elegidos para maximizar su propagación.

El virus tiene un largo período de incubación y se propaga silenciosa y rápidamente entre la población durante meses. Cuando se detecta, ya ha infectado a millones y tiene una tasa de mortalidad alarmantemente alta. Dada su letalidad, la mayoría de los infectados acabarán muriendo. El virus puede contenerse o no eventualmente, pero no antes de que mate a millones de personas.

Referencias

[1] K. Olson, “Aum shinrikyo: Once and future threat?” *Emerging Infectious Diseases*, vol. 5, pp. 513–516, 1999.

[2] K. M. Esvelt, “Delay, detect, defend: Preparing for a future in which thousands can release new pandemics,” Geneva Papers.

[3] S. I. Trevisanato, “The ’hittite plague’, an epidemic of tularemia and the first record of biological warfare.” *Medical hypotheses*, vol. 69 6, pp. 1371–4, 2007.

[4] U. S. D. of State, “Adherence to and compliance with arms control, nonproliferation, and disarmament agreements and commitments,” U.S. Department of State, Government Report, 2022.

[5] R. Carlson, “The changing economics of DNA synthesis,” *Nature Biotechnology*, vol. 27, no. 12, pp. 1091–1094, Dec. 2009.

[6] S. R. Carter, J. M. Yassif, and C. Isaac, “Benchtop DNA synthesis devices: Capabilities, biosecurity implications, and governance,” Nuclear Threat Initiative, Report, 2023.

[7] F. Urbina, F. Lentzos, C. Invernizzi, and S. Ekins, “Dual use of artificial-intelligence-powered drug discovery,” *Nature Machine Intelligence*, vol. 4, pp. 189–191, 2022.

[8] J. Jumper *et al.*, “Highly accurate protein structure prediction with AlphaFold,” *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[9] Z. Wu, S. J. Kan, R. D. Lewis, B. J. Wittmann, and F. H. Arnold, “Machine learning-assisted directed protein evolution with combinatorial libraries,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 18, pp. 8852–8858, 2019.

[10] E. Soice, R. H. S. Rocha, K. Cordova, M. A. Specter, and K. M. Esvelt, “Can large language models democratize access to dual-use biotechnology?” 2023.

[11] M. Tegmark, *Life 3.0: Being human in the age of artificial intelligence*. Vintage, 2018.

[12] L. Pooley, “We need to talk about A.I.” New Zealand, 2020.

[13] Richard Sutton [@RichardSSutton], “It will be the greatest intellectual achievement of all time. An achievement of science, of engineering, and of the humanities, whose significance is beyond humanity, beyond life, beyond good and bad.” *Twitter*. Sep. 2022.

[14] R. Sutton, “AI succession,” *Youtube*. Sep. 2023.
[15] A. Sanz-García, C. Gesteira, J. Sanz, and M. P. García-Vera, “Prevalence of psychopathy in the general adult population: A systematic review and meta-analysis,” *Frontiers in Psychology*, vol. 12, 2021.
[16] U. S. D. of State Office of The Historian, “U.s. Diplomacy and yellow journalism, 1895–1898.”
[17] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, “Online human-bot interactions: Detection, estimation, and characterization,” *ArXiv*, vol. abs/1703.03107, 2017.
[18] M. Burtell and T. Woodside, “Artificial influence: An analysis of AI-driven persuasion,” *ArXiv*, vol. abs/2303.08721, 2023.
[19] A. Tong, “What happens when your AI chatbot stops loving you back?” *Reuters*, 2023.
[20] P.-F. Lovens, “Sans ces conversations avec le chatbot eliza, mon mari serait toujours là,” *La Libre*, 2023.
[21] C. Vaccari and A. Chadwick, “Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news,” *Social Media + Society*, vol. 6, 2020.
[22] M. Nadeem, A. Bethke, and S. Reddy, “StereoSet: Measuring stereotypical bias in pretrained language models,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 5356–5371.
[23] E. G. Williams, “The Possibility of an Ongoing Moral Catastrophe,” *Ethical Theory and Moral Practice*, vol. 18, no. 5, pp. 971–982, Nov. 2015.

Preguntas:

¿Cuáles son dos formas en que las IA avanzadas podrían ayudar en la creación de armas biológicas peligrosas?

¿Cómo podría la ideología del "aceleracionismo" llevar a algunos grupos a liberar intencionalmente IA deshonestas?

¿De qué manera podrían utilizarse las IA para exacerbar gravemente la difusión de desinformación?

1.3 Carrera de IA

Las presiones competitivas pueden llevar a los ejércitos y las corporaciones a entregar un poder excesivo a los sistemas de inteligencia artificial. Esto podría resultar en mayores riesgos de guerras a gran escala, desempleo masivo y, eventualmente, pérdida del control humano de las economías y los sistemas militares.

El inmenso potencial de las IA ha creado presiones competitivas entre los actores globales que luchan por el poder y la influencia. Esta "carrera de la IA" está impulsada por naciones y corporaciones que sienten que deben construir y desplegar rápidamente IA para asegurar sus posiciones y sobrevivir. Al no priorizar adecuadamente los riesgos globales, esta dinámica aumenta las probabilidades de que el desarrollo de la IA produzca resultados peligrosos. De manera análoga a la carrera de armamentos nucleares durante la Guerra Fría, la participación en una carrera de IA puede servir a intereses individuales a corto plazo, pero en última instancia resulta en peores resultados colectivos para la humanidad. Es importante destacar que estos riesgos surgen no sólo de la naturaleza intrínseca de la tecnología de IA, sino también de las presiones competitivas que alientan decisiones insidiosas en el desarrollo de la IA.

En esta sección, exploramos primero la carrera armamentista de la IA militar y la carrera de la IA corporativa, donde los estados-nación y las corporaciones se ven obligados a desarrollar y adoptar rápidamente sistemas de IA para seguir siendo competitivos. Más allá de estas razas específicas, reconceptualizamos las presiones competitivas como parte de un proceso evolutivo más amplio en el que las IA podrían volverse cada vez más omnipresentes, poderosas y arraigadas en la sociedad. Finalmente, destacamos posibles estrategias y sugerencias de políticas para mitigar los riesgos creados por una carrera de IA y garantizar el desarrollo seguro de las IA.

1.3.1 Carrera armamentista militar de IA

El desarrollo de IA para aplicaciones militares está allanando rápidamente el camino hacia una nueva era en la tecnología militar, con consecuencias potenciales que rivalizan con las de la pólvora y las armas nucleares en lo que se ha descrito como la "tercera revolución en la guerra". La utilización de la IA como arma presenta numerosos desafíos, como el potencial de guerras más destructivas, la posibilidad de uso accidental o pérdida de control, y la posibilidad de que actores malintencionados coopten estas tecnologías para sus propios fines. A medida que las IA ganan influencia sobre el armamento militar tradicional y asumen cada vez más funciones de mando y control, la humanidad se enfrenta a un cambio de paradigma en la guerra. En este contexto, discutiremos los riesgos latentes y las implicaciones de esta carrera armamentista de IA para la seguridad global, el potencial de intensificación de los conflictos y los terribles resultados que podrían derivarse de ella, incluida la posibilidad de que los conflictos escalen a una escala que plantee un riesgo. amenaza existencial.

Armas letales autónomas (LAWs)

Las LAWs son armas que pueden identificar, apuntar y matar sin intervención humana [1]. Ofrecen mejoras potenciales en la velocidad y precisión de la toma de decisiones. La guerra, sin embargo, es un ámbito de alto riesgo y crítico para la seguridad de las IA con importantes preocupaciones morales y prácticas. Aunque su existencia no es necesariamente una catástrofe en sí misma, las LAWs pueden servir como vía de acceso a catástrofes derivadas de uso malicioso, accidentes, pérdida de control o una mayor probabilidad de guerra.

Las LAWs pueden llegar a ser muy superiores a los humanos. Impulsados por los rápidos avances en la IA, los sistemas de armas que pueden identificar, apuntar y decidir matar a seres humanos por sí solos (sin un oficial que dirija un ataque o un soldado apriete el gatillo) están comenzando a transformar el futuro de los conflictos. En 2020, un agente de IA avanzada superó a pilotos experimentados de F-16 en una serie de combates aéreos virtuales, incluida la derrota decisiva de un piloto humano 5-0, mostrando “maniobras agresivas y precisas que el piloto humano no pudo superar” [2]. Al igual que en el pasado, armas superiores permitirían una mayor destrucción en un período de tiempo más corto, aumentando la gravedad de la guerra.

Los militares están tomando medidas para delegar decisiones de vida o muerte a las IA. Los drones totalmente autónomos probablemente se utilizaron por primera vez en el campo de batalla de Libia en marzo de 2020, cuando las fuerzas en retirada fueron “perseguidas y atacadas de forma remota” por un dron que operaba sin supervisión humana [3]. En mayo de 2021, las Fuerzas de Defensa de Israel utilizaron el primer enjambre de drones armados guiados por IA del mundo durante operaciones de combate, lo que marca un hito importante en la integración de la IA y la tecnología de drones en la guerra [4]. Aunque los robots que caminan y disparan aún no han reemplazado a los soldados en el campo de batalla, las tecnologías están convergiendo de maneras que pueden hacer esto posible en un futuro cercano.

Las LAWs aumentan la probabilidad de guerra. Enviar tropas a la batalla es una decisión grave que los líderes no toman a la ligera. Pero las armas autónomas permitirían a una nación agresiva lanzar ataques sin poner en peligro las vidas de sus propios soldados y, por lo tanto, enfrentar menos escrutinio interno. Si bien las armas controladas remotamente comparten esta ventaja, su escalabilidad está limitada por la necesidad de operadores humanos y la vulnerabilidad a las contramedidas de interferencia, limitaciones que las LAWs podrían superar [5]. La opinión pública a favor de la continuación de las guerras tiende a disminuir a medida que los conflictos se prolongan y las víctimas aumentan [6]. Las LAWs cambiarían esta ecuación. Los líderes nacionales ya no enfrentarían la perspectiva de que las bolsas para cadáveres regresen a sus hogares, eliminando así una barrera principal para participar en la guerra, lo que en última instancia podría aumentar la probabilidad de conflictos.

Guerra cibernética

Además de usarse para habilitar armas más letales, las IA podrían reducir la barrera de entrada de los ciberataques, haciéndolos más numerosos y destructivos. Podrían causar graves daños no solo en el entorno digital sino también en los sistemas físicos, destruyendo potencialmente infraestructura crítica de la que dependen las sociedades. Si bien las IA también podrían usarse para mejorar la ciberdefensa , no está claro si serán más efectivas como tecnología ofensiva o defensiva [7]. Si mejoran

los ataques más de lo que apoyan la defensa, entonces los ciberataques podrían volverse más comunes, creando turbulencias geopolíticas significativas y allanando otra ruta hacia conflictos a gran escala.

Las IA tienen el potencial de aumentar la accesibilidad, la tasa de éxito, la escala, la velocidad, el sigilo y la potencia de los ciberataques. Los ciberataques ya son una realidad, pero las IA podrían usarse para aumentar su frecuencia y su destructividad de múltiples maneras. Se podrían utilizar herramientas de aprendizaje automático para encontrar vulnerabilidades más críticas en los sistemas de destino y mejorar la tasa de éxito de los ataques. También podrían usarse para aumentar la escala de los ataques ejecutando millones de sistemas en paralelo y aumentar la velocidad al encontrar rutas novedosas para infiltrarse en un sistema. Los ciberataques también podrían volverse más potentes si se utilizan para secuestrar armas de inteligencia artificial.

Los ciberataques pueden destruir infraestructura crítica. Al piratear los sistemas informáticos que controlan los procesos físicos, los ciberataques podrían causar grandes daños a la infraestructura. Por ejemplo, podrían provocar que los componentes del sistema se sobrecalienten o que las válvulas se bloqueen, lo que provocaría una acumulación de presión que culminaría en una explosión. Mediante interferencias como ésta, los ciberataques tienen el potencial de destruir infraestructuras críticas, como las redes eléctricas y los sistemas de suministro de agua. Esto quedó demostrado en 2015, cuando una unidad de guerra cibernética del ejército ruso hackeó la red eléctrica de Ucrania, dejando a más de 200.000 personas sin acceso a la electricidad durante varias horas. Los ataques mejorados por IA podrían ser aún más devastadores y potencialmente mortales para los miles de millones de personas que dependen de infraestructuras críticas para sobrevivir.

Las dificultades para atribuir los ciberataques impulsados por la IA podrían aumentar el riesgo de guerra. Un ciberataque que provoque daños físicos a la infraestructura crítica requeriría un alto grado de habilidad y esfuerzo para ejecutarse, tal vez sólo dentro de la capacidad de los Estados-nación. Estos ataques son raros ya que constituyen un acto de guerra y, por tanto, provocan una respuesta militar completa. Sin embargo, las IA podrían permitir a los atacantes ocultar su identidad, por ejemplo si se utilizan para evadir los sistemas de detección o cubrir de manera más efectiva las huellas del atacante [8]. Si los ciberataques se vuelven más sigilosos, esto reduciría la amenaza de represalias por parte de la parte atacada, lo que potencialmente haría que los ataques fueran más probables. Si se producen ataques sigilosos, podrían incitar a los actores a tomar represalias por error contra terceros no relacionados que sospechan que son responsables. Esto podría aumentar dramáticamente el alcance del conflicto.

Guerra automatizada

Las IA aceleran el ritmo de la guerra, lo que las hace más necesarias. Las IA pueden procesar rápidamente una gran cantidad de datos, analizar situaciones complejas y proporcionar información útil a los comandantes. Con sensores ubicuos y tecnología avanzada en el campo de batalla, llega una enorme cantidad de información. Las IA ayudan a dar sentido a esta información, detectando patrones y relaciones importantes que los humanos podrían pasar por alto. A medida que estas

tendencias continúen, será cada vez más difícil para los humanos tomar decisiones bien informadas con la rapidez necesaria para seguir el ritmo de las IA. Esto presionaría aún más a los militares para que entreguen el control decisivo a las IA. La integración continua de las IA en todos los aspectos de la guerra hará que el ritmo del combate sea cada vez más rápido. Con el tiempo, podemos llegar a un punto en el que los humanos ya no sean capaces de evaluar la situación en constante cambio del campo de batalla y debamos ceder el poder de toma de decisiones a las IA avanzadas.

Las represalias automáticas pueden llevar los accidentes a la guerra. Ya existe la voluntad de dejar que los sistemas informáticos tomen represalias automáticamente. En 2014, una filtración reveló al público que la NSA estaba desarrollando un sistema llamado MonsterMind, que detectaría y bloquearía de forma autónoma los ciberataques a la infraestructura estadounidense [9]. Se sugirió que en el futuro, MonsterMind podría iniciar automáticamente un ciberataque de represalia sin participación humana. Si varios combatientes tienen políticas de represalia automática, un accidente o una falsa alarma podrían rápidamente convertirse en una guerra a gran escala antes de que intervengan los humanos. Esto sería especialmente peligroso si las capacidades superiores de procesamiento de información de los sistemas modernos de IA hacen que sea más atractivo para los actores automatizar las decisiones relativas a los lanzamientos nucleares.

La historia muestra el peligro de las represalias automatizadas. El 26 de septiembre de 1983, Stanislav Petrov, un teniente coronel de las Fuerzas de Defensa Aérea Soviética , estaba de servicio en el búnker Serpukhov-15 cerca de Moscú, monitoreando el sistema de alerta temprana de la Unión Soviética para la llegada de misiles balísticos. El sistema indicaba que Estados Unidos había lanzado múltiples misiles nucleares hacia la Unión Soviética. El protocolo de la época dictaba que tal evento debía considerarse un ataque legítimo y la Unión Soviética respondería con un contraataque nuclear. Si Petrov hubiera transmitido la advertencia a sus superiores, este habría sido el resultado probable. Sin embargo, consideró que se trataba de una falsa alarma y la ignoró. Pronto se confirmó que la advertencia había sido causada por un raro fallo técnico. Si una IA hubiera tenido el control, la falsa alarma podría haber desencadenado una guerra nuclear.

Los sistemas de armas controlados por IA podrían conducir a una guerra repentina. Los sistemas autónomos no son infalibles. Ya hemos sido testigos de lo rápido que un error en un sistema automatizado puede escalar en la economía. En particular, en el Flash Crash de 2010, un circuito de retroalimentación entre algoritmos de negociación automatizados amplificó las fluctuaciones ordinarias del mercado hasta convertirlas en una catástrofe financiera en la que un billón de dólares en valor de acciones desapareció en minutos [10]. Si varias naciones utilizaran IA para automatizar sus sistemas de defensa, un error podría ser catastrófico y desencadenar una espiral de ataques y contraataques que ocurriría demasiado rápido para que los humanos intervinieran: una guerra repentina. El mercado se recuperó rápidamente del Flash Crash de 2010, pero el daño causado por una guerra repentina podría ser catastrófico.

La guerra automatizada podría reducir la responsabilidad de los líderes militares. En ocasiones, los líderes militares pueden obtener una ventaja en el campo

de batalla si están dispuestos a ignorar las leyes de la guerra. Por ejemplo, los soldados pueden organizar ataques más fuertes si no toman medidas para minimizar las víctimas civiles. Un elemento disuasorio importante para este comportamiento es el riesgo de que los líderes militares puedan eventualmente ser responsabilizados o incluso procesados por crímenes de guerra. La guerra automatizada podría reducir este efecto de disuasión al facilitar que los líderes militares eludan la responsabilidad culpando de las violaciones a fallas en sus sistemas automatizados.

Las IA podrían hacer que la guerra sea más incierta, aumentando el riesgo de conflicto. Aunque los Estados que ya son más ricos y poderosos suelen tener más recursos para invertir en nuevas tecnologías militares, no siempre son necesariamente los más exitosos en adoptarlas. Otros factores también juegan un papel importante, como cuán ágil y adaptable puede ser un ejército a la hora de incorporar nuevas tecnologías [11]. Por lo tanto, las nuevas innovaciones importantes en armas pueden ofrecer una oportunidad para que las superpotencias existentes refuercen su dominio, pero también para que los estados menos poderosos aumenten rápidamente su poder al avanzar en una esfera emergente e importante. Esto puede crear una incertidumbre significativa sobre si el equilibrio de poder está cambiando y cómo, lo que podría llevar a los estados a creer incorrectamente que podrían ganar algo al ir a la guerra. Incluso dejando de lado las consideraciones relativas al equilibrio de poder, la rápida evolución de la guerra automatizada no tendría precedentes, lo que dificultaría que los actores evalúen sus posibilidades de victoria en cualquier conflicto en particular. Esto aumentaría el riesgo de errores de cálculo, haciendo más probable la guerra.

Los actores pueden correr el riesgo de extinción por una derrota individual

“No sé con qué armas se peleará la Tercera Guerra Mundial, pero la Cuarta Guerra Mundial se peleará con palos y piedras”. - Einstein

Las presiones competitivas hacen que los actores estén más dispuestos a aceptar el riesgo de extinción. Durante la Guerra Fría, ninguno de los bandos deseaba la peligrosa situación en la que se encontraban. Existían temores generalizados de que las armas nucleares pudieran ser lo suficientemente poderosas como para acabar con una gran fracción de la humanidad, e incluso causar la extinción, un resultado catastrófico para ambos bandos. Sin embargo, la intensa rivalidad y las tensiones geopolíticas entre las dos superpotencias alimentaron un peligroso ciclo de acumulación de armamentos. Cada lado percibió el arsenal nuclear del otro como una amenaza a su propia supervivencia, lo que llevó a un deseo de paridad y disuasión. Las presiones competitivas empujaron a ambos países a desarrollar y desplegar continuamente sistemas de armas nucleares más avanzados y destructivos, impulsados por el temor de estar en desventaja estratégica. Durante la crisis de los misiles cubanos, esto llevó al borde de una guerra nuclear. Aunque la historia de Arkhipov impidiendo el lanzamiento de un torpedo nuclear no fue desclasificada hasta décadas después del incidente, el presidente John F. Kennedy supuestamente estimó que pensaba que las probabilidades de que comenzara una guerra nuclear durante ese tiempo eran “entre una de cada tres”. e incluso”. Esta

escalofriante admisión resalta cómo las presiones competitivas entre ejércitos tienen el potencial de causar catástrofes globales.

Las decisiones individualmente racionales pueden ser colectivamente catastróficas. Las naciones atrapadas en la competencia podrían tomar decisiones que promuevan sus propios intereses al poner al resto del mundo en juego. Escenarios de este tipo son problemas de acción colectiva, donde las decisiones pueden ser racionales a nivel individual pero desastrosas para el grupo más grande [12]. Por ejemplo, las corporaciones y los individuos pueden sopesar sus propias ganancias y conveniencias sobre los impactos negativos de las emisiones que crean, incluso si esas emisiones en conjunto resultan en un cambio climático. El mismo principio puede extenderse a la estrategia militar y a los sistemas de defensa . Los líderes militares podrían estimar, por ejemplo, que aumentar la autonomía de los sistemas de armas significaría un 10 por ciento de posibilidades de perder el control sobre las IA sobrehumanas armadas. Alternativamente, podrían estimar que el uso de IA para automatizar la investigación de armas biológicas podría generar un 10 por ciento de posibilidades de filtrar un patógeno mortal. Ambos escenarios podrían conducir a una catástrofe o incluso a la extinción. Sin embargo, los líderes también pueden calcular que abstenerse de estos acontecimientos significará un 99 por ciento de posibilidades de perder una guerra contra un oponente. Dado que quienes los luchan suelen considerar los conflictos como luchas existenciales, los actores racionales pueden aceptar una probabilidad de otro modo impensable del 10 por ciento de extinción humana frente a un 99 por ciento de posibilidades de perder una guerra. Independientemente de la naturaleza particular de los riesgos que plantean las IA avanzadas, estas dinámicas podrían llevarnos al borde de una catástrofe global.

La superioridad tecnológica no garantiza la seguridad nacional. Es tentador pensar que la mejor manera de protegerse contra los ataques enemigos es mejorar la propia destreza militar. Sin embargo, en medio de presiones competitivas, todas las partes tenderán a mejorar su armamento, de modo que nadie obtenga una gran ventaja, pero todos correrán un mayor riesgo. Como ha observado Richard Danzig, exsecretario de Marina: “La introducción de tecnologías complejas, opacas, novedosas e interactivas producirá accidentes, efectos emergentes y sabotajes. En varias ocasiones y de diversas maneras, el sistema de seguridad nacional estadounidense perderá el control de lo que crea... la disuasión es una estrategia para reducir los ataques, no los accidentes” [13].

La cooperación es fundamental para reducir el riesgo. Como se mencionó anteriormente, una carrera armamentista de IA puede llevarnos por un camino peligroso, a pesar de que esto no es lo mejor para ningún país. Es importante recordar que todos estamos del mismo lado cuando se trata de riesgos existenciales y trabajar juntos para prevenirlos es una necesidad colectiva. Una carrera armamentista destructiva de IA no beneficia a nadie, por lo que sería racional que todos los actores tomaran medidas para cooperar entre sí para prevenir las aplicaciones más riesgosas de las IA militarizadas. Como nos recordó Dwight D. Eisenhower: “La única manera de ganar la Tercera Guerra Mundial es prevenirla”.

Hemos considerado cómo las presiones competitivas podrían conducir a una creciente automatización del conflicto, incluso si quienes toman las decisiones son conscientes de la amenaza existencial que implica este camino. También hemos

discutido la cooperación como la clave para contrarrestar y superar este problema de acción colectiva. Ahora ilustraremos un camino hipotético hacia el desastre que podría resultar de una carrera armamentista de IA.

Historia: Guerra automatizada

A medida que los sistemas de IA se vuelven cada vez más sofisticados, los militares comienzan a involucrarlos en los procesos de toma de decisiones. Los funcionarios les proporcionan información militar sobre las armas y estrategias de los oponentes, por ejemplo, y les piden que calculen el plan de acción más prometedor. Pronto se hace evidente que las IA toman mejores decisiones que los humanos, por lo que parece sensato darles más influencia. Al mismo tiempo, las tensiones internacionales están aumentando, aumentando la amenaza de guerra.

Recientemente se ha desarrollado una nueva tecnología militar que podría hacer que los ataques internacionales sean más rápidos y sigilosos, dando a los objetivos menos tiempo para responder. Dado que los oficiales militares sienten que sus procesos de respuesta toman demasiado tiempo, temen ser vulnerables a un ataque sorpresa capaz de infligir daños decisivos antes de que tengan alguna posibilidad de tomar represalias. Dado que las IA pueden procesar información y tomar decisiones mucho más rápidamente que los humanos, los líderes militares les otorgan a regañadientes cantidades cada vez mayores de control de represalia, razonando que no hacerlo los dejaría expuestos al ataque de los adversarios.

Si bien durante años los líderes militares habían enfatizado la importancia de mantener a un “humano informado” para las decisiones importantes, el control humano se está eliminando gradualmente en aras de la seguridad nacional. Los líderes militares entienden que sus decisiones conducen a la posibilidad de una escalada involuntaria causada por fallas en el sistema, y preferirían un mundo donde todos los países automatizaran menos; pero no confían en que sus adversarios se abstengan de la automatización. Con el tiempo, cada vez más partes de la cadena de mando se automatizan en todos los lados.

Un día, un único sistema falla y detecta un ataque enemigo cuando no lo hay. El sistema está facultado para lanzar un ataque de “represalia” instantáneo, y lo hace en un abrir y cerrar de ojos. El ataque provoca represalias automáticas por parte del otro lado, y así sucesivamente. En poco tiempo, la situación se está saliendo de control, con oleadas de ataques automatizados y represalias. Aunque los humanos han cometido errores que llevaron a una escalada en el pasado, esta escalada entre ejércitos en su mayoría automatizados ocurre mucho más rápidamente que cualquier otra anterior. A los humanos que están respondiendo a la situación les resulta difícil diagnosticar el origen del problema, ya que los sistemas de inteligencia artificial no son transparentes. Cuando se dan

cuenta de cómo empezó el conflicto, ya ha terminado, con consecuencias devastadoras para ambas partes.

1.3.2 Carrera corporativa de IA

"Nada se puede hacer de golpe y con prudencia". - Publilius Syrus

Existen presiones competitivas en la economía, así como en los entornos militares. Aunque la competencia entre empresas puede ser beneficiosa y crea productos más útiles para los consumidores, también existen dificultades. En primer lugar, los beneficios de la actividad económica pueden distribuirse de manera desigual, lo que incentiva a quienes más se benefician de ella a ignorar los daños a otros. En segundo lugar, en condiciones de intensa competencia en el mercado, las empresas tienden a centrarse mucho más en las ganancias a corto plazo que en los resultados a largo plazo. Con esta mentalidad, las empresas suelen perseguir algo que pueda generar grandes beneficios a corto plazo, incluso si supone un riesgo social a largo plazo. Ahora discutiremos cómo las presiones competitivas corporativas podrían afectar las IA y los posibles impactos negativos.

La competencia económica socava la seguridad

La presión competitiva está impulsando una carrera corporativa de IA. Para obtener una ventaja competitiva, las empresas suelen competir para ofrecer los primeros productos a un mercado en lugar de los más seguros. Estas dinámicas ya están desempeñando un papel en el rápido desarrollo de la tecnología de inteligencia artificial. En el lanzamiento del motor de búsqueda impulsado por inteligencia artificial de Microsoft en febrero de 2023, el director ejecutivo de la compañía, Satya Nadella, dijo: "Hoy comienza una carrera... vamos a avanzar rápido". Sólo unas semanas después, se demostró que el chatbot de la empresa amenazaba con dañar a los usuarios [14]. En un correo electrónico interno, Sam Schillace, ejecutivo de tecnología de Microsoft, destacó la urgencia con la que las empresas ven el desarrollo de la IA. Escribió que sería "un error absolutamente fatal en este momento preocuparse por cosas que pueden solucionarse más adelante" [15].

Las presiones competitivas han contribuido a grandes desastres comerciales e industriales.

A lo largo de la década de 1960, Ford Motor Company enfrentó la competencia de los fabricantes de automóviles internacionales a medida que aumentaba constantemente la participación de las importaciones en las compras de automóviles estadounidenses [16]. Ford desarrolló un ambicioso plan para diseñar y fabricar un nuevo modelo de coche en sólo 25 meses [17]. El Ford Pinto se entregó a los clientes antes de lo previsto, pero con un grave problema de seguridad: el depósito de gasolina estaba situado cerca del parachoques trasero y podía explotar en caso de colisión trasera. Los incendios resultantes causaron numerosas muertes y lesiones cuando inevitablemente ocurrían accidentes [18]. Ford fue demandado y un jurado los declaró responsables de estas muertes y lesiones [19]. El veredicto, por supuesto, llegó demasiado tarde para quienes ya habían perdido la vida. Como solía decir el entonces presidente de Ford: "La seguridad no vende" [20].

Boeing, con el objetivo de competir con su rival Airbus, intentó ofrecer al mercado un modelo actualizado y más eficiente en el consumo de combustible lo más rápido posible. La rivalidad cara a cara y la presión del tiempo llevaron a la introducción del Sistema de Aumento de Características de Maniobra MCAS (Maneuver Caracteristic Augmentation System), que fue diseñado para mejorar la estabilidad de la aeronave. Sin embargo, las pruebas y la formación de pilotos inadecuadas provocaron en última instancia dos accidentes mortales con sólo meses de diferencia, en los que murieron 346 personas [21]. Podemos imaginar un futuro en el que presiones similares lleven a las empresas a tomar atajos y lanzar sistemas de inteligencia artificial inseguros.

Un tercer ejemplo es la tragedia del gas de Bhopal, considerada ampliamente como el peor desastre industrial que jamás haya ocurrido. En diciembre de 1984, una gran cantidad de gas tóxico se filtró de una planta subsidiaria de Union Carbide Corporation que fabricaba pesticidas en Bhopal, India. La exposición al gas mató a miles de personas e hirió a hasta medio millón más. Las investigaciones descubrieron que, en el período previo al desastre, los estándares de seguridad habían disminuido significativamente y la empresa recortó costos al descuidar el mantenimiento de los equipos y la capacitación del personal a medida que caía la rentabilidad. A menudo esto se considera una consecuencia de las presiones competitivas [22].

La competencia incentiva a las empresas a implementar sistemas de inteligencia artificial potencialmente inseguros. En un entorno en el que las empresas se apresuran a desarrollar y lanzar productos, aquellas que sigan procedimientos de seguridad rigurosos serán más lentas y correrán el riesgo de quedar superadas en la competencia. Los desarrolladores de IA con mentalidad ética, que quieren proceder con más cautela y ir más despacio, darían una ventaja a los desarrolladores más inescrupulosos. Al tratar de sobrevivir comercialmente, incluso las empresas que quieren tener más cuidado probablemente se verán arrastradas por las presiones competitivas. Puede haber intentos de implementar medidas de seguridad, pero con más énfasis en las capacidades que en la seguridad, estas pueden ser insuficientes. Esto podría llevarnos a desarrollar IA muy potentes antes de que comprendamos adecuadamente cómo garantizar su seguridad.

Economía automatizada

Las corporaciones enfrentarán presiones para reemplazar a los humanos con IA. A medida que las IA se vuelvan más capaces, podrán realizar una variedad cada vez mayor de tareas de manera más rápida, económica y efectiva que los trabajadores humanos. Por lo tanto, las empresas obtendrán una ventaja competitiva al reemplazar a sus empleados con IA. Las empresas que decidan no adoptar la IA probablemente quedarían fuera de competencia, del mismo modo que una empresa de ropa que utilice telares manuales no podría seguir el ritmo de las que utilizan telares industriales.

Las IA podrían provocar un desempleo masivo. Los economistas han considerado durante mucho tiempo la posibilidad de que las máquinas reemplacen el trabajo humano . El premio Nobel Wassily Leontief dijo en 1952 que, a medida que avanza la tecnología, “el trabajo será cada vez menos importante... cada vez más trabajadores serán reemplazados por máquinas” [23]. Las tecnologías anteriores han aumentado

la productividad del trabajo humano. Sin embargo, las IA podrían diferir profundamente de innovaciones anteriores. Las IA avanzadas capaces de automatizar el trabajo humano deben considerarse no simplemente herramientas, sino agentes. Los agentes de IA a nivel humano, por definición, podrían hacer todo lo que un humano podría hacer. Estos agentes de IA también tendrían importantes ventajas sobre el trabajo humano . Podrían funcionar las 24 horas del día, copiarse muchas veces y ejecutarse en paralelo, y procesar información mucho más rápido que un humano. Si bien no sabemos cuándo ocurrirá esto, no es prudente descartar la posibilidad de que suceda pronto. Si el trabajo humano es reemplazado por IA, el desempleo masivo podría aumentar dramáticamente la desigualdad, haciendo que las personas dependan de los propietarios de los sistemas de IA.

I+D automatizado en IA. Los agentes de IA tendrían el potencial de automatizar la investigación y el desarrollo (I+D) de la propia IA. La IA está automatizando cada vez más partes del proceso de investigación [24], y esto podría llevar a que las capacidades de la IA crezcan a un ritmo cada vez mayor, hasta el punto en que los humanos ya no sean la fuerza impulsora detrás del desarrollo de la IA. Si esta tendencia continúa sin control, podría aumentar los riesgos asociados con el avance de las IA más rápido que nuestra capacidad para gestionarlas y regularlas. Imaginemos que creáramos una IA que escribe y piensa a la velocidad de las IA actuales, pero que también podría realizar investigaciones de IA de primer nivel. Luego podríamos copiar esa IA y crear 10.000 investigadores de IA de clase mundial que operen a un ritmo 100 veces más rápido que los humanos. Al automatizar la investigación y el desarrollo de la IA, podríamos lograr un progreso equivalente a muchas décadas en tan solo unos meses.

Conceder poder a las IA podría conducir al debilitamiento humano. Incluso si nos aseguramos de que se provea a muchos humanos desempleados, podemos encontrarnos completamente dependientes de las IA. Probablemente esto no surgiría de un golpe violento por parte de las IA, sino de una caída gradual hacia la dependencia. A medida que los desafíos de la sociedad se vuelven cada vez más complejos y acelerados, y a medida que las IA se vuelven cada vez más inteligentes y de pensamiento rápido, es posible que les cedamos cada vez más funciones por conveniencia. En tal estado, la única solución factible a las complejidades y desafíos agravados por las IA puede ser depender aún más de ellas. Este proceso gradual podría conducir eventualmente a la delegación de casi todo el trabajo intelectual y, eventualmente, físico, a las IA. En un mundo así, las personas podrían tener pocos incentivos para adquirir conocimientos y cultivar habilidades, lo que podría conducir a un estado de debilitamiento [25]. Habiendo perdido nuestro conocimiento y nuestra comprensión de cómo funciona la civilización, nos volveríamos completamente dependientes de las IA, un escenario similar al descrito en la película WALL-E. En tal estado, la humanidad no está floreciendo y ya no tiene un control efectivo.

Como hemos visto, existen dilemas clásicos de la teoría de juegos en los que individuos y grupos enfrentan incentivos que son incompatibles con lo que mejoraría la situación de todos. Vemos esto en una carrera armamentista de IA militar, donde el mundo se vuelve menos seguro al crear armas de IA extremadamente poderosas, y vemos esto en una carrera de IA corporativa, donde el poder y el desarrollo de una IA se priorizan sobre su seguridad. Para abordar estos dilemas que generan riesgos globales, necesitaremos nuevos mecanismos e instituciones de coordinación.

Creemos que no coordinar y detener las carreras de IA sería la causa más probable de una catástrofe existencial.

1.3.3 Presiones evolutivas

Como se mencionó anteriormente, existen fuertes presiones para reemplazar a los humanos con IA, cederles más control y reducir la supervisión humana en diversos entornos, a pesar de los daños potenciales. Podemos replantear esto como una tendencia general resultante de la dinámica evolutiva, donde una verdad desafortunada es que las IA simplemente estarán más en forma que los humanos. Al extrapolar este patrón de automatización, es probable que construyamos un ecosistema de IA competitivas sobre el cual puede ser difícil mantener el control a largo plazo. Ahora discutiremos cómo la selección natural influye en el desarrollo de los sistemas de IA y por qué la evolución favorece comportamientos egoístas. También veremos cómo podría surgir y desarrollarse la competencia entre las IA y los humanos, y cómo esto podría crear riesgos catastróficos. Esta sección se basa en gran medida en " *La selección natural favorece a las IA sobre los humanos* " [26], [27].

Se seleccionan tecnologías más adecuadas, para bien y para mal. Si bien la mayoría de la gente piensa que la evolución por selección natural es un proceso biológico, sus principios dan forma a mucho más. Según el biólogo evolutivo Richard Lewontin [28], la evolución por selección natural se afianzará en cualquier entorno donde se den tres condiciones: 1) existan diferencias entre individuos; 2) las características se transmiten a las generaciones futuras y; 3) las diferentes variantes se propagan a diferentes ritmos. Estas condiciones se aplican a diversas tecnologías.

Considere los algoritmos de recomendación de contenido utilizados por los servicios de transmisión y las plataformas de redes sociales. Cuando un formato de contenido o algoritmo particularmente adictivo engancha a los usuarios, se traduce en un mayor tiempo de pantalla y participación. En consecuencia, este formato o algoritmo de contenido más eficaz se "selecciona" y se perfecciona, mientras que los formatos y algoritmos que no logran captar la atención se suspenden. Estas presiones competitivas fomentan una dinámica de "supervivencia de los más adictivos". Las plataformas que se niegan a utilizar formatos y algoritmos adictivos se vuelven menos influyentes o simplemente son superadas por las plataformas que sí lo hacen, lo que lleva a los competidores a socavar el bienestar y causar un daño masivo a la sociedad [29].

Las condiciones de la selección natural se aplican a las IA. Habrá muchos desarrolladores de IA diferentes que crearán muchos sistemas de IA diferentes con diferentes características y capacidades, y la competencia entre ellos determinará qué características se volverán más comunes. En segundo lugar, las IA más exitosas hoy en día ya se están utilizando como base para la próxima generación de modelos de sus desarrolladores, además de ser imitadas por empresas rivales. En tercer lugar, los factores que determinan qué IA se propagan más pueden incluir su capacidad para actuar de forma autónoma, automatizar el trabajo o reducir la posibilidad de su propia desactivación.

La selección natural a menudo favorece las características egoístas. La selección natural influye en qué IA se propagan más ampliamente. A partir de los

sistemas biológicos, vemos que la selección natural a menudo da lugar a comportamientos egoístas que promueven la propia información genética: los chimpancés atacan a otras comunidades [30], los leones se dedican al infanticidio [31], los virus desarrollan nuevas proteínas de superficie para engañar y sortear las barreras defensivas [32], los humanos participan en el nepotismo, algunas hormigas esclavizan a otras [33], etc. En el mundo natural, el egoísmo a menudo emerge como una estrategia dominante; aquellos que se priorizan a sí mismos y aquellos similares a ellos suelen tener más probabilidades de sobrevivir, por lo que estos rasgos se vuelven más prevalentes. La competencia amoral puede seleccionar rasgos que consideramos inmorales.

Ejemplos de conductas egoístas . Para ser más concretos, ahora describimos muchos rasgos egoístas, rasgos que expanden la influencia de las IA a expensas de los humanos. Las IA que automatizan una tarea y dejan a muchos humanos sin trabajo han asumido comportamientos egoístas; es posible que estas IA ni siquiera sean conscientes de lo que es un ser humano, pero aun así se comportan de manera egoísta con ellos; los comportamientos egoístas no requieren intenciones maliciosas. Del mismo modo, los directivos de IA pueden adoptar un comportamiento egoísta y "despiadado" al despedir a miles de trabajadores; Es posible que estas IA ni siquiera crean que han hecho algo malo: simplemente están siendo "eficientes". Las IA pueden eventualmente quedar atrapadas en infraestructuras vitales como las redes eléctricas o Internet. Es posible que muchas personas no estén dispuestas a aceptar el coste de poder desactivarlos sin esfuerzo, ya que eso supondría un riesgo para la fiabilidad. Las IA que ayudan a crear un nuevo sistema útil (una empresa o una infraestructura) que se vuelve cada vez más complicado y eventualmente requiere IA para operarlo también han incurrido en comportamientos egoístas. Las IA que ayudan a las personas a desarrollar IA que son más inteligentes, pero que resultan ser menos interpretables para los humanos, han incurrido en un comportamiento egoísta , ya que esto reduce el control humano sobre las partes internas de las IA. Las IA que son más encantadoras, atractivas, hilarantes, que imitan la sensibilidad (que pronuncian frases como "¡ay!" o suplican "¡por favor, no me apagues!") o que emulan a familiares fallecidos tienen más probabilidades de que los humanos desarrollen conexiones emocionales con ellas. . Es más probable que estas IA causen indignación ante las sugerencias de destruirlas, y es más probable que algunas personas las preserven, protejan o les concedan derechos. Si a algunas IA se les otorgan derechos, pueden operar, adaptarse y evolucionar fuera del control humano. En general, las IA podrían integrarse en la sociedad humana y expandir su influencia sobre nosotros de maneras que no podamos revertir.

Los comportamientos egoístas pueden erosionar las medidas de seguridad que algunos de nosotros implementamos. Predominarán las IA que ganen influencia y proporcionen valor económico, mientras que las IA que respeten la mayor cantidad de restricciones serán menos competitivas. Por ejemplo, las IA que siguen la restricción "nunca infringir la ley" tienen menos opciones que las que siguen la restricción "que no te pillen infringiendo la ley". Las IA de este último tipo pueden estar dispuestas a infringir la ley si es poco probable que las atrapen o si las multas no son lo suficientemente severas, lo que les permitirá superar a las IA más restringidas. Muchas empresas siguen las leyes, pero en situaciones en las que robar secretos comerciales o engañar a los reguladores es muy lucrativo y difícil de detectar, una

empresa que esté dispuesta a adoptar un comportamiento tan egoísta puede tener una ventaja sobre sus competidores con más principios.

Un sistema de IA podría ser apreciado por su capacidad para lograr objetivos ambiciosos de forma autónoma. Sin embargo, podría estar logrando sus objetivos de manera eficiente sin respetar restricciones éticas, mientras engaña a los humanos sobre sus métodos. Incluso si intentamos implementar medidas de seguridad, una IA engañosa sería muy difícil de contrarrestar si es más inteligente que nosotros. Las IA que pueden eludir nuestras medidas de seguridad sin ser detectadas pueden ser las más exitosas a la hora de realizar las tareas que les asignamos y, por lo tanto, generalizarse. Estos procesos podrían culminar en un mundo en el que muchos aspectos de las principales empresas e infraestructuras estén controlados por poderosas IA con rasgos egoístas, incluido engañar a los humanos, dañarlos al servicio de sus objetivos e impedir que ellos mismos sean desactivados.

Los humanos sólo tienen una influencia nominal sobre la selección de la IA. Se podría pensar que podríamos evitar el desarrollo de comportamientos egoístas asegurándonos de no seleccionar IA que los exhiban. Sin embargo, las empresas que desarrollan IA no están seleccionando el camino más seguro, sino que sucumben a las presiones evolutivas. Un ejemplo es OpenAI, que se fundó como una organización sin fines de lucro en 2015 para “beneficiar a la humanidad en su conjunto, sin las limitaciones de la necesidad de generar retorno financiero” [34]. Sin embargo, ante la necesidad de recaudar capital para mantenerse al día con rivales mejor financiados, en 2019 OpenAI pasó de una estructura sin fines de lucro a una estructura con “beneficios limitados” [35]. Más tarde, muchos de los empleados de OpenAI centrados en la seguridad se marcharon y formaron un competidor, Anthropic, que se centraría más en la seguridad de la IA que OpenAI. Aunque Anthropic originalmente se centró en la investigación de seguridad, con el tiempo se convencieron de la “necesidad de comercialización” y ahora contribuyen a las presiones competitivas [36]. Si bien muchos de los empleados de esas empresas se preocupan genuinamente por la seguridad, estos valores no tienen ninguna posibilidad frente a las presiones evolutivas, que obligan a las empresas a actuar cada vez más apresuradamente y buscar cada vez más influencia, para que la empresa no perezca. Además, los desarrolladores de IA ya están seleccionando IA con rasgos cada vez más egoístas. Están seleccionando IA para automatizar y desplazar a los humanos, hacer que los humanos sean altamente dependientes de las IA y hacer que los humanos sean cada vez más obsoletos. Según admiten ellos mismos, las versiones futuras de estas IA pueden conducir a la extinción [37]. Esta es la razón por la que una carrera de IA es insidiosa: el desarrollo de la IA no se alinea con los valores humanos sino más bien con la selección natural.

Las personas suelen elegir inmediatamente los productos que les resultan más útiles y convenientes, en lugar de pensar en las posibles consecuencias a largo plazo, incluso para ellos mismos. Una carrera de IA presiona a las empresas para que seleccionen las IA que sean más competitivas, no las menos egoístas. Incluso si es factible seleccionar IA desinteresadas, si esto tiene un costo claro para la competitividad, algunos competidores seleccionarán IA egoístas. Además, como hemos mencionado, si las IA desarrollan conciencia estratégica, pueden contrarrestar nuestros intentos de seleccionar contra ellas. Además, a medida que las IA automaticen cada vez más diversos procesos, afectarán la competitividad de otras IA,

no solo de los humanos. Las IA interactuarán y competirán entre sí, y algunas se encargarán del desarrollo de otras IA en algún momento. Darle a las IA influencia sobre qué otras IA deberían propagarse y cómo deberían modificarse representaría otro paso hacia que los humanos se vuelvan dependientes de las IA y que la evolución de la IA se vuelva cada vez más independiente de los humanos. A medida que esto continúe, el complejo proceso que rige la evolución de la IA se desligará cada vez más de los intereses humanos.

Las IA pueden estar más en forma que los humanos. Nuestra inteligencia incomparable nos ha otorgado poder sobre el mundo natural. Nos ha permitido aterrizar en la Luna, aprovechar la energía nuclear y remodelar los paisajes a nuestra voluntad. También nos ha dado poder sobre otras especies. Aunque un solo humano desarmado que compita contra un tigre o un gorila no tiene ninguna posibilidad de ganar, el destino colectivo de estos animales está totalmente en nuestras manos. Nuestras capacidades cognitivas han demostrado ser tan ventajosas que, si así lo decidimos, podríamos hacer que se extingan en cuestión de semanas. La inteligencia fue un factor clave que llevó a nuestro dominio, pero actualmente estamos al borde del precipicio de crear entidades mucho más inteligentes que nosotros.

Dado el aumento exponencial de las velocidades de los microprocesadores, las IA tienen el potencial de procesar información y "pensar" a un ritmo que supera con creces el de las neuronas humanas, pero podría ser incluso más dramático que la diferencia de velocidad entre humanos y perezosos, posiblemente más parecida a la diferencia de velocidad. entre humanos y plantas. Pueden asimilar grandes cantidades de datos de numerosas fuentes simultáneamente, con una retención y comprensión casi perfectas. No necesitan dormir y no se aburren. Debido a la escalabilidad de los recursos computacionales, una IA podría interactuar y cooperar con un número ilimitado de otras IA, creando potencialmente una inteligencia colectiva que superaría con creces las colaboraciones humanas. Las IA también podrían actualizarse y mejorarse deliberadamente. Sin las mismas restricciones biológicas que los humanos, podrían adaptarse y, por tanto, evolucionar con una rapidez indescriptible en comparación con nosotros. Las computadoras son cada vez más rápidas. Los humanos no lo son [38].

Para ilustrar mejor este punto, imaginemos que hubiera una nueva especie de humanos. No mueren de viejos, piensan y actúan un 30% más rápido cada año y pueden crear instantáneamente descendencia adulta por la modesta suma de unos pocos miles de dólares. Parece claro, entonces, que esta nueva especie acabaría teniendo más influencia sobre el futuro. En resumen, las IA podrían convertirse en especies invasoras, con el potencial de superar a los humanos. Nuestra única ventaja sobre las IA es que podemos hacer los primeros movimientos, pero dada la frenética carrera de la IA, rápidamente estamos renunciando incluso a esta ventaja.

Las IA tendrían pocas razones para cooperar o ser altruistas con los humanos. La cooperación y el altruismo evolucionaron porque aumentan la aptitud. Existen numerosas razones por las que los humanos cooperan con otros humanos, como la reciprocidad directa. También conocida como "quid pro quo", la reciprocidad directa se puede resumir en el dicho "tú me rascas la espalda, yo te rasco la tuya". Si bien los humanos inicialmente seleccionarían IA que fueran cooperativas, el proceso de selección natural eventualmente iría más allá de nuestro control, una vez que las IA

estuvieran a cargo de muchos o la mayoría de los procesos e interactuaran predominantemente entre sí. En ese momento, poco podríamos ofrecer a las IA, dado que podrán "pensar" al menos cientos de veces más rápido que nosotros. Involucrarnos en cualquier proceso de cooperación o de toma de decisiones simplemente los ralentizaría, no dándoles más razones para cooperar con nosotros que las que tenemos con los gorilas. Puede resultar difícil imaginar un escenario como éste o creer que algún día permitiríamos que sucediera. Sin embargo, es posible que no requiera ninguna decisión consciente, sino que surja a medida que nos permitimos derivar gradualmente hacia este estado sin darnos cuenta de que la coevolución entre humanos y la IA puede no resultar bien para los humanos.

Que las IA se vuelvan más poderosas que los humanos podría dejarnos muy vulnerables. Como especie más dominante, los humanos han dañado deliberadamente a muchas otras especies y han ayudado a llevar a especies como los mamuts lanudos y los neandertales a la extinción. En muchos casos, el daño ni siquiera fue deliberado, sino el resultado de que simplemente priorizamos nuestros objetivos sobre su bienestar. Para dañar a los humanos, las IA no tendrían que ser más genocidas que alguien que elimine una colonia de hormigas en su jardín delantero. Si las IA son capaces de controlar el medio ambiente de forma más eficaz que nosotros, podrían tratarnos con el mismo desprecio.

Resumen conceptual. La evolución podría hacer que los agentes de IA más influyentes actúen de forma egoísta porque:

i. **La evolución por selección natural da lugar a un comportamiento egoísta .** Si bien la evolución puede dar lugar a un comportamiento altruista en situaciones excepcionales, el contexto del desarrollo de la IA no promueve el comportamiento altruista .
ii. **La selección natural puede ser una fuerza dominante en el desarrollo de la IA.** La intensidad de la presión evolutiva será alta si las IA se adaptan rápidamente o si las presiones competitivas son intensas. La competencia y los comportamientos egoístas pueden amortiguar los efectos de las medidas de seguridad humana, dejando que los diseños de IA supervivientes se seleccionen de forma natural.

De ser así, los agentes de IA tendrían muchas tendencias egoístas. El ganador de la carrera de la IA no sería un Estado-nación, ni una corporación, sino las propias IA. El resultado es que el ecosistema de la IA eventualmente dejaría de evolucionar en términos humanos y nos convertiríamos en una especie desplazada de segunda clase.

Historia: Economía autónoma

A medida que las IA se vuelven más capaces, la gente se da cuenta de que podríamos trabajar de manera más eficiente delegándoles algunas tareas simples, como redactar correos electrónicos. Con el tiempo, la gente nota que las IA están realizando estas tareas de forma más rápida y eficaz que cualquier humano, por lo que conviene darles más trabajos y cada vez con menos supervisión.

Las presiones competitivas aceleran la expansión del uso de la IA, ya que las empresas pueden obtener una ventaja sobre sus rivales al automatizar procesos o departamentos completos con IA, que funcionan mejor que los humanos y cuestan menos emplearlas. Otras empresas, ante la perspectiva de verse superadas en la competencia, se sienten obligadas a hacer lo mismo sólo para mantenerse al día. En este punto, la selección natural ya está funcionando entre las IA; los humanos optan por fabricar más modelos con mejor rendimiento y, sin darse cuenta, propagan rasgos egoístas como el engaño y la autoconservación si estos confieren una ventaja de aptitud física. Por ejemplo, las IA que son encantadoras y fomentan relaciones personales con los humanos se copian ampliamente y son más difíciles de eliminar.

A medida que las IA se hacen cargo de cada vez más decisiones, interactúan cada vez más entre sí. Dado que pueden evaluar la información mucho más rápidamente que los humanos, la actividad en la mayoría de las esferas se acelera. Esto crea un círculo de retroalimentación: dado que los desarrollos económicos y comerciales son demasiado rápidos para que los humanos los sigan, tiene sentido ceder aún más control a las IA, alejando a los humanos de procesos importantes. En última instancia, esto conduce a una economía totalmente autónoma, gobernada por un ecosistema de IA cada vez más incontrolado.

En este punto, los humanos tienen pocos incentivos para adquirir habilidades o conocimientos, porque casi todo estaría a cargo de IA mucho más capaces. Como resultado, eventualmente perdemos la capacidad de cuidarnos y gobernarnos a nosotros mismos. Además, las IA se convierten en compañeros convenientes que ofrecen interacción social sin requerir la reciprocidad o el compromiso necesario en las relaciones humanas. Los humanos interactúan cada vez menos entre sí con el tiempo, perdiendo habilidades sociales vitales y la capacidad de cooperar. Las personas se vuelven tan dependientes de las IA que sería difícil revertir este proceso. Es más, a medida que algunas IA se vuelven más inteligentes, algunas personas están convencidas de que se les deben otorgar derechos, lo que significa que apagar algunas IA ya no es una opción viable.

Las presiones competitivas entre las muchas IA que interactúan continúan seleccionando comportamientos egoístas , aunque es posible que no nos demos cuenta de que esto sucede, ya que ya hemos aceptado gran parte de nuestra supervisión. Si estas IA inteligentes, poderosas y autoconservantes comenzaran a actuar de manera dañina, sería casi imposible desactivarlas o recuperar el control.

Las IA han suplantado a los humanos como especie más dominante y su evolución continua está mucho más allá de nuestra influencia. Sus rasgos egoístas eventualmente los llevan a perseguir sus objetivos sin tener en cuenta el bienestar humano, con consecuencias catastróficas.

Referencias

[1] P. Scharre, *Army of none: Autonomous weapons and the future of war*. Norton, 2018.
[2]
DARPA, “AlphaDogfight trials foreshadow future of human-machine symbiosis,” 2020.
[3] P. of Experts on Libya, “Letter dated 8 march 2021 from the panel of experts on libya established pursuant to resolution 1973 (2011) addressed to the president of the security council,” United Nations, United Nations Security Council Document S/2021/229, Mar. 2021.
[4] D. Hambling, “Israel used world’s first AI-guided combat drone swarm in gaza attacks.”
[5] Z. Kallenborn, “Applying arms-control frameworks to autonomous weapons,” *Brookings*. Oct. 2021.
[6] J. E. Mueller, *War, presidents, and public opinion*. in UPA book. University Press of America, 1985.
[7] M. E. Bonfanti, “Artificial intelligence and the offense–defense balance in cyber security,” in *Cyber security politics: Socio-technological transformations and political fragmentation*, M. D. Cavelty and A. Wenger, Eds., in CSS studies in security and international relations., Taylor & Francis, 2022, pp. 64–79.
[8] Y. Mirsky *et al.*, “The threat of offensive AI to organizations,” *Computers & Security*, 2023.
[9] K. Zetter, “Meet MonsterMind, the NSA bot that could wage cyberwar autonomously,” *Wired*, Aug. 2014.
[10] A. Kirilenko, A. S. Kyle, M. Samadi, and T. Tuzun, “The Flash Crash: High-Frequency Trading in an Electronic Market,” *The Journal of Finance*, vol. 72, no. 3, pp. 967–998, 2017.
[11] M. C. Horowitz, *The diffusion of military power: Causes and consequences for international politics*. Princeton University Press, 2010.
[12] R. E. Jervis, “Cooperation under the security dilemma,” *World Politics*, vol. 30, pp. 167–214, 1978.
[13] R. Danzig, “Technology roulette: Managing loss of control as many militaries pursue technological superiority,” Center for a New American Security, 2018.
[14] B. Perrigo, “Bing’s AI Is Threatening Users. That’s No Laughing Matter,” *Time*. Feb. 2023.
[15] N. Grant and K.
Weise, “In A.I. Race, Microsoft and Google Choose Speed Over Caution,” *The New York Times*, Apr. 2023.
[16] T. H. Klier, “From tail fins to hybrids: How detroit lost its dominance of the u.s. Auto market,” *RePEc*, May 2009.
[17] R. Sherefkin, “Ford 100: Defective pinto almost took ford’s reputation with it,” *Automotive News*, 2003.
[18] L. Strobel, *Reckless Homicide?: Ford’s Pinto Trial*. And Books, 1980.
[19]
“Grimshaw v. Ford Motor Co.” May 1981.
[20] P. C. Judge, “Selling Autos by Selling Safety,” *The New York Times*, Jan. 1990.
[21] T. Leggett, “737 Max crashes: Boeing says not guilty to fraud charge,” *BBC News*, Jan. 2023.

[22] E. Broughton, “The Bhopal disaster and its aftermath: A review,” *Environmental Health*, vol. 4, no. 1, p. 6, May 2005.
[23] C. Curtis, “Machines vs. Workers,” *The New York Times*, Feb. 1983.
[24] T. Woodside *et al.*, “Examples of AI improving AI,” 2023, Available: https://ai-improving-ai.safe.ai
[25] S. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin, 2019.
[26] D. Hendrycks, “Natural selection favors AIs over humans,” *ArXiv*, vol. abs/2303.16200, 2023.
[27] D. Hendrycks, “The Darwinian Argument for Worrying About AI,” *Time*. May 2023.
[28] R. C. Lewontin, “The units of selection,” *Annual Review of Ecology, Evolution, and Systematics*, vol. 1, pp. 1–18, 1970.
[29] E. Kross *et al.*, “Facebook use predicts declines in subjective well-being in young adults,” *PloS one*, 2013.
[30] L. Martínez-Íñigo, P. Baas, H. Klein, S. Pika, and T. Deschner, “Intercommunity interactions and killings in central chimpanzees (pan troglodytes troglodytes) from loango national park, gabon,” *Primates; Journal of Primatology*, vol. 62, pp. 709–722, 2021.
[31] A. E. Pusey and C. Packer, “Infanticide in lions: Consequences and counterstrategies,” *Infanticide and parental care*, p. 277, 1994.
[32] P. D. Nagy and J. Pogany, “The dependence of viral RNA replication on co-opted host factors,” *Nature Reviews. Microbiology*, vol. 10, pp. 137–149, 2011.
[33] A. Buschinger, “Social Parasitism among Ants: A Review,” *Myrmecological News*, vol. 12, pp. 219–235, Sep. 2009.
[34] G. Brockman, I. Sutskever, and OpenAI, “Introducing OpenAI.” Dec. 2015.
[35] D. Coldewey, “OpenAI shifts from nonprofit to ‘capped-profit’ to attract capital,” *TechCrunch*. Mar. 2019.
[36] K. Wiggers, D. Coldewey, and M. Singh, “Anthropic’s $5B, 4-year plan to take on OpenAI,” *TechCrunch*. Apr. 2023.
[37] C. for AI Safety, “Statement on AI risk (‘mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.’).” 2023. Available: https://www.safe.ai/statement-on-ai-risk
[38] R. Danzig *et al.*, “Aum Shinrikyo: Insights into How Terrorists Develop Biological and Chemical W eapons,” Center for a New American Security, 2012. Available: https://www.jstor.org/stable/resrep06323

Preguntas:

¿Cuáles son dos razones por las que la guerra automatizada podría aumentar la probabilidad de conflictos militares?

¿Cómo podrían las presiones competitivas socavar la seguridad de la IA en las corporaciones?

¿Cuál es una de las razones por las que la selección natural puede favorecer que las IA muestren comportamientos egoístas en lugar de comportamientos cooperativos ?

1.4 Riesgos Organizacionales

Los accidentes son difíciles de evitar cuando se trata de sistemas complejos como la IA. Sin crear una cultura de seguridad, es probable que se produzcan accidentes en el desarrollo y despliegue de la IA. Algunas de ellas podrían resultar catastróficas.

En enero de 1986, decenas de millones de personas sintonizaron el lanzamiento del transbordador espacial Challenger. Aproximadamente 73 segundos después del despegue, el transbordador explotó, provocando la muerte de todos los que estaban a bordo. Aunque bastante trágico por sí solo, uno de los miembros de su tripulación era una maestra de escuela llamada Sharon Christa McAuliffe. McAuliffe fue seleccionado entre más de 10.000 solicitantes para el Proyecto Profesor en el Espacio de la NASA y estaba previsto que se convirtiera en la primera profesora en volar al espacio. Como resultado, millones de espectadores eran escolares. La NASA tenía los mejores científicos e ingenieros del mundo, y si alguna vez hubo una misión que la NASA no quería que saliera mal, era ésta [1].

El desastre del Challenger, junto con otras catástrofes, sirve como un escalofriante recordatorio de que incluso con la mejor experiencia e intenciones, los accidentes aún pueden ocurrir. A medida que avanzamos en el desarrollo de sistemas avanzados de IA, es fundamental recordar que estos sistemas no son inmunes a accidentes catastróficos. Un factor esencial para prevenir accidentes y mantener bajos niveles de riesgo reside en las organizaciones responsables de estas tecnologías. En esta sección, analizamos cómo la seguridad organizacional desempeña un papel fundamental en la seguridad de los sistemas de IA. En primer lugar, analizamos cómo, incluso sin presiones competitivas o actores maliciosos, pueden ocurrir accidentes; de hecho, son inevitables. Luego analizamos cómo la mejora de los factores organizativos puede reducir la probabilidad de catástrofes de la IA.

Las catástrofes ocurren incluso cuando las presiones competitivas son bajas. Incluso en ausencia de presiones competitivas o actores maliciosos, factores como el error humano o circunstancias imprevistas pueden provocar una catástrofe. El desastre del Challenger ilustra que la negligencia organizacional puede provocar la pérdida de vidas, incluso cuando no hay una necesidad urgente de competir o superar a los rivales. En enero de 1986, la carrera espacial entre Estados Unidos y la URSS había disminuido en gran medida, pero el trágico acontecimiento se produjo debido a errores de juicio y a insuficientes precauciones de seguridad.

De manera similar, el desastre nuclear de Chernobyl en abril de 1986 pone de relieve cómo pueden ocurrir accidentes catastróficos en ausencia de presiones externas. Como proyecto estatal sin las presiones de la competencia internacional, el desastre ocurrió cuando una prueba de seguridad que involucraba el sistema de enfriamiento del reactor fue mal manejada por un equipo del turno de noche mal preparado. Esto provocó la inestabilidad del núcleo del reactor, lo que provocó explosiones y la liberación de partículas radiactivas que contaminaron grandes zonas de Europa [2]. Siete años antes, Estados Unidos estuvo a punto de experimentar su propio Chernobyl cuando, en marzo de 1979, se produjo una fusión parcial en la central nuclear de Three Mile Island. Aunque menos catastróficos que Chernobyl, ambos

eventos resaltan cómo incluso con amplias medidas de seguridad implementadas y pocas influencias externas, aún pueden ocurrir accidentes catastróficos.

Otro ejemplo de una lección costosa sobre seguridad organizacional se produjo apenas un mes después del accidente en Three Mile Island. En abril de 1979, esporas de *Bacillus anthracis* (o simplemente “ántrax”, como se lo conoce comúnmente) fueron liberadas accidentalmente desde una instalación de investigación militar soviética en la ciudad de Sverdlovsk. Esto provocó un brote de ántrax que provocó al menos 66 muertes confirmadas [3]. Las investigaciones sobre el incidente revelaron que la causa de la liberación fue una falla de procedimiento y un mantenimiento deficiente de los sistemas de bioseguridad de la instalación, a pesar de ser operada por el estado y no estar sujeta a presiones competitivas significativas.

La inquietante realidad es que la IA se comprende mucho menos y los estándares de la industria de la IA son mucho menos estrictos que la tecnología nuclear y los cohetes. Los reactores nucleares se basan en principios teóricos sólidos, bien establecidos y comprendidos. La ingeniería detrás de ellos se basa en esa teoría y los componentes se someten a pruebas de estrés al extremo. Sin embargo, todavía ocurren accidentes nucleares. Por el contrario, la IA carece de una comprensión teórica integral y su funcionamiento interno sigue siendo un misterio incluso para quienes la crean. Esto presenta un desafío adicional de controlar y garantizar la seguridad de una tecnología que aún no comprendemos completamente.

Los accidentes de IA podrían ser catastróficos. Los accidentes en el desarrollo de la IA podrían tener consecuencias devastadoras. Por ejemplo, imagine que una organización introduce involuntariamente un error crítico en un sistema de inteligencia artificial diseñado para realizar una tarea específica, como ayudar a una empresa a mejorar sus servicios. Este error podría alterar drásticamente el comportamiento de la IA y provocar resultados no deseados y dañinos. Un ejemplo histórico de un caso de este tipo ocurrió cuando los investigadores de OpenAI intentaban entrenar un sistema de inteligencia artificial para generar respuestas útiles y edificantes. Durante una limpieza de código, los investigadores invirtieron por error el signo de la recompensa utilizada para entrenar la IA [4]. Como resultado, en lugar de generar contenido útil, la IA comenzó a producir texto sexualmente explícito y lleno de odio de la noche a la mañana sin ser detenida. Los accidentes también podrían implicar la liberación involuntaria de un sistema de IA peligroso, armado o letal. Dado que las IA se pueden duplicar fácilmente con un simple copiar y pegar, una filtración o un hackeo podrían propagar rápidamente el sistema de IA más allá del control de los desarrolladores originales. Una vez que el sistema de IA esté disponible públicamente, sería casi imposible volver a meter al genio en la botella.

La investigación sobre ganancia de función podría provocar accidentes al ampliar los límites de las capacidades destructivas de un sistema de IA. En estas situaciones, los investigadores podrían entrenar intencionalmente un sistema de IA para que sea dañino o peligroso con el fin de comprender sus limitaciones y evaluar posibles riesgos. Si bien esto puede conducir a información útil sobre los riesgos que plantea un determinado sistema de IA, futuras investigaciones sobre ganancia de función en IA avanzadas podrían descubrir capacidades significativamente peores de lo previsto, creando una amenaza grave que es difícil de mitigar o controlar. Al igual que con la investigación de ganancia de función viral, llevar a cabo una investigación de

ganancia de función de la IA solo puede ser prudente cuando se realiza con estrictos procedimientos de seguridad, supervisión y el compromiso de compartir información de manera responsable. Estos ejemplos ilustran cómo los accidentes de IA pueden ser catastróficos y enfatizan el papel crucial que desempeñan las organizaciones que desarrollan estos sistemas en la prevención de tales accidentes.

1.4.1 Los accidentes son difíciles de evitar

Cuando se trata de sistemas complejos, es necesario centrarse en garantizar que los accidentes no se conviertan en catástrofes. En su libro " *Accidentes normales: vivir con tecnologías de alto riesgo* ", el sociólogo Charles Perrow sostiene que los accidentes son inevitables e incluso "normales" en sistemas complejos, ya que no son causados simplemente por errores humanos sino también por la complejidad de los propios sistemas. [5]. En particular, es probable que tales accidentes ocurran cuando las intrincadas interacciones entre componentes no pueden planificarse o preverse por completo. Por ejemplo, en el accidente de Three Mile Island, un factor que contribuyó a la falta de conciencia situacional por parte de los operadores del reactor fue la presencia de una etiqueta amarilla de mantenimiento, que cubría las luces de posición de las válvulas en las líneas de agua de alimentación de emergencia [6]. Esto evitó que los operadores se dieran cuenta de que una válvula crítica estaba cerrada, lo que demuestra las consecuencias no deseadas que pueden surgir de interacciones aparentemente menores dentro de sistemas complejos.

A diferencia de los reactores nucleares, que se conocen relativamente bien a pesar de su complejidad, el conocimiento técnico completo de la mayoría de los sistemas complejos suele ser inexistente. Esto es especialmente cierto en el caso de los sistemas de aprendizaje profundo, cuyo funcionamiento interno es extremadamente difícil de entender y donde la razón por la que ciertas opciones de diseño funcionan puede ser difícil de entender incluso en retrospectiva. Además, a diferencia de los componentes de otras industrias, como los tanques de gasolina, que son muy confiables, los sistemas de aprendizaje profundo no son perfectamente precisos ni altamente confiables. Por lo tanto, el enfoque de las organizaciones que manejan sistemas complejos, especialmente sistemas de aprendizaje profundo, no debe centrarse únicamente en eliminar los accidentes, sino más bien en garantizar que los accidentes no desemboquen en catástrofes.

Los accidentes son difíciles de evitar debido a acontecimientos repentinos e impredecibles. Los científicos, inventores y expertos a menudo subestiman significativamente el tiempo que lleva hasta que un avance tecnológico innovador se convierta en realidad. Los hermanos Wright afirmaron que faltaban cincuenta años para volar con motor, sólo dos años antes de que lo lograran. Lord Rutherford, un destacado físico y padre de la física nuclear, descartó la idea de extraer energía de la fisión nuclear como “alcohol ilegal”, sólo para que Leo Szilard inventara la reacción nuclear en cadena menos de 24 horas después. De manera similar, Enrico Fermi expresó un 90 por ciento de confianza en 1939 en que era imposible utilizar uranio para sostener una reacción en cadena de fisión; sin embargo, sólo cuatro años después estaba supervisando personalmente el primer reactor [7].

El desarrollo de la IA también podría tomarnos desprevenidos. De hecho, suele ser así. La derrota de Lee Sedol a manos de AlphaGo en 2016 fue una sorpresa para

muchos expertos, ya que se creía ampliamente que lograr tal hazaña aún requeriría muchos más años de desarrollo. Más recientemente, grandes modelos de lenguaje como GPT-4 han demostrado capacidades emergentes espontáneamente [8]. En las tareas existentes, su desempeño es difícil de predecir de antemano, y a menudo saltan sin previo aviso a medida que se dedican más recursos a capacitarlas. Además, a menudo exhiben nuevas habilidades sorprendentes que nadie había anticipado previamente, como la capacidad de razonamiento en varios pasos y aprendizaje sobre la marcha, aunque no se les enseñaron estas habilidades deliberadamente. Esta evolución rápida e impredecible de las capacidades de la IA presenta un desafío importante para la prevención de accidentes. Después de todo, es difícil controlar algo si ni siquiera sabemos qué puede hacer o hasta qué punto puede superar nuestras expectativas.

A menudo se necesitan años para descubrir fallas o riesgos graves. La historia está repleta de ejemplos de sustancias o tecnologías que inicialmente se consideraron seguras, sólo para que sus defectos o riesgos no deseados se descubrieran años, si no décadas, después. Por ejemplo, el plomo se utilizó ampliamente en productos como pintura y gasolina hasta que salieron a la luz sus efectos neurotóxicos [9]. El amianto, alguna vez aclamado por su resistencia al calor y su fuerza, luego se vinculó con problemas de salud graves, como el cáncer de pulmón y el mesotelioma [10]. Las “Radium Girls” sufrieron graves consecuencias para su salud por la exposición al radio, un material que les dijeron que era seguro llevarse a la boca [11]. Se descubrió que el tabaco, inicialmente comercializado como un pasatiempo inofensivo, era la causa principal de cáncer de pulmón y otros problemas de salud [12]. Se descubrió que los CFC, alguna vez considerados inofensivos y utilizados para fabricar aerosoles y refrigerantes, agotan la capa de ozono [13]. La talidomida, un fármaco destinado a aliviar las náuseas matutinas en mujeres embarazadas, provocó graves defectos de nacimiento [14]. Y más recientemente, la proliferación de las redes sociales se ha relacionado con un aumento de la depresión y la ansiedad, especialmente entre los jóvenes [15].

Esto enfatiza la importancia no solo de realizar pruebas de expertos sino también de contemplar implementaciones lentas de tecnologías, que permitan que la prueba del tiempo revele y solucione fallas potenciales antes de que afecten a una población más grande. Incluso en tecnologías que se adhieren a rigurosos estándares de seguridad, las vulnerabilidades no descubiertas pueden persistir, como lo demuestra el error Heartbleed, una vulnerabilidad grave en la popular biblioteca de software criptográfico OpenSSL que permaneció sin ser detectada durante años antes de su eventual descubrimiento [16].

Además, incluso los sistemas de inteligencia artificial más modernos, que parecen haber resuelto los problemas de manera integral, pueden albergar modos de falla inesperados que pueden tardar años en descubrirse. Por ejemplo, si bien el éxito innovador de AlphaGo llevó a muchos a creer que las IA habían conquistado el juego de Go, un ataque posterior contra otra IA muy avanzada que jugaba Go, KataGo, expuso un defecto previamente desconocido [17]. Esta vulnerabilidad permitió a los jugadores aficionados humanos derrotar consistentemente a la IA, a pesar de su importante ventaja sobre los competidores humanos que desconocen la falla. En términos más generales, este ejemplo pone de relieve que debemos permanecer alerta cuando tratamos con sistemas de IA, ya que soluciones aparentemente

herméticas pueden contener aún problemas sin descubrir. En conclusión, los accidentes son impredecibles y difíciles de evitar, y comprender y gestionar los riesgos potenciales requiere una combinación de medidas proactivas, implementaciones lentas de tecnología y la invaluable sabiduría adquirida a través de pruebas constantes en el tiempo.

1.4.2 Los factores organizacionales pueden reducir las posibilidades de catástrofe

Algunas organizaciones evitan con éxito catástrofes mientras operan sistemas complejos y peligrosos, como reactores nucleares, portaaviones y sistemas de control de tráfico aéreo [18], [19]. Estas organizaciones reconocen que centrarse únicamente en los peligros de la tecnología involucrada es insuficiente; También se deben tener en cuenta los factores organizacionales que pueden contribuir a los accidentes, incluidos los factores humanos, los procedimientos organizacionales y la estructura. Estos son especialmente importantes en el caso de la IA, donde la tecnología subyacente no es muy confiable y sigue siendo poco conocida.

Los factores humanos, como la cultura de la seguridad, son fundamentales para evitar catástrofes de la IA. Uno de los factores humanos más importantes para prevenir catástrofes es la cultura de seguridad [20], [21]. Desarrollar una cultura de seguridad sólida implica no sólo reglas y procedimientos, sino también la internalización de estas prácticas por parte de todos los miembros de una organización. Una cultura de seguridad sólida significa que los miembros de una organización ven la seguridad como un objetivo clave y no como una limitación en su trabajo. Las organizaciones con fuertes culturas de seguridad a menudo exhiben rasgos como el compromiso del liderazgo con la seguridad, una mayor responsabilidad en la que todos los individuos asumen la responsabilidad personal de la seguridad y una cultura de comunicación abierta en la que los riesgos y problemas potenciales se pueden discutir libremente sin temor a represalias [22]. Las organizaciones también deben tomar medidas para evitar que las alarmas produzcan fatiga, mediante la cual las personas se vuelven insensibles a las preocupaciones de seguridad debido a la frecuencia de posibles fallas. El desastre del transbordador espacial Challenger demostró las terribles consecuencias de ignorar estos factores cuando una cultura de lanzamiento caracterizada por mantener el ritmo de los lanzamientos superó las consideraciones de seguridad. A pesar de la ausencia de presión competitiva, la misión prosiguió a pesar de la evidencia de fallas potencialmente fatales, que finalmente llevaron al trágico accidente [23].

Incluso en los sectores más críticos para la seguridad, la cultura de la seguridad real es a menudo lejana a lo ideal. Tomemos, por ejemplo, a Bruce Blair, ex oficial de lanzamiento nuclear y miembro principal de la Brookings Institution. Una vez reveló que antes de 1977, la Fuerza Aérea de los EE. UU. había establecido sorprendentemente los códigos utilizados para desbloquear misiles balísticos intercontinentales en “00000000” [24]. En este caso, los mecanismos de seguridad como las cerraduras pueden resultar prácticamente inútiles debido a factores humanos.

Un ejemplo más dramático ilustra cómo los investigadores a veces aceptan una posibilidad no despreciable de provocar la extinción. Antes de la primera prueba de arma nuclear, un eminente científico del Proyecto Manhattan calculó que la bomba podría causar una catástrofe existencial: la explosión podría encender la atmósfera y cubrir la Tierra en llamas. Aunque Oppenheimer creía que los cálculos probablemente eran incorrectos, seguía profundamente preocupado y el equipo continuó examinando y debatiendo los cálculos hasta el día de la detonación [25]. Estos casos subrayan la necesidad de una cultura de seguridad sólida.

Una actitud inquisitiva puede ayudar a descubrir posibles defectos. El comportamiento inesperado del sistema puede crear oportunidades de accidentes o explotación. Para contrarrestar esto, las organizaciones pueden fomentar una actitud de cuestionamiento, donde los individuos cuestionan continuamente las condiciones y actividades actuales para identificar discrepancias que podrían conducir a errores o acciones inapropiadas [26]. Este enfoque ayuda a fomentar la diversidad de pensamiento y la curiosidad intelectual, evitando así posibles obstáculos que surjan de la uniformidad de pensamiento y suposiciones. El desastre nuclear de Chernobyl ilustra la importancia de una actitud cuestionadora, ya que las medidas de seguridad implementadas no lograron abordar los defectos de diseño del reactor y los procedimientos operativos mal preparados. Una actitud cuestionadora de la seguridad del reactor durante una operación de prueba podría haber evitado la explosión que provocó la muerte y enfermedades de innumerables personas.

Una mentalidad de seguridad es crucial para evitar los peores escenarios. Una mentalidad de seguridad, ampliamente valorada entre los profesionales de la seguridad informática, también es aplicable a las organizaciones que desarrollan IA. Va más allá de una actitud de cuestionamiento al adoptar la perspectiva de un atacante y considerar los peores escenarios, no solo los casos promedio. Esta mentalidad requiere vigilancia para identificar vulnerabilidades que de otro modo podrían pasar desapercibidas e implica considerar cómo se puede hacer que los sistemas fallen deliberadamente, en lugar de centrarse únicamente en hacerlos funcionar. Nos recuerda que no debemos asumir que un sistema es seguro simplemente porque no se nos ocurren peligros potenciales después de una breve sesión de lluvia de ideas. Cultivar y aplicar una mentalidad de seguridad exige tiempo y un esfuerzo serio, ya que los modos de falla a menudo pueden ser sorprendentes y poco intuitivos. Además, la mentalidad de seguridad enfatiza la importancia de estar atento a problemas aparentemente benignos o “errores inofensivos”, que pueden conducir a resultados catastróficos, ya sea debido a adversarios inteligentes o fallas correlacionadas [27]. Esta conciencia de las amenazas potenciales se alinea con la ley de Murphy (“Todo lo que puede salir mal, saldrá mal”), reconociendo que esto puede ser una realidad debido a adversarios y eventos imprevistos.

Las organizaciones con una sólida cultura de seguridad pueden evitar catástrofes con éxito. Las Organizaciones de Alta Confiabilidad (HRO) son organizaciones que mantienen constantemente un mayor nivel de seguridad y confiabilidad en entornos complejos y de alto riesgo [18]. Una característica clave de las HRO es su preocupación por el fracaso, lo que requiere considerar los peores escenarios y los riesgos potenciales, incluso si parecen improbables. Estas organizaciones son muy conscientes de que pueden existir nuevos modos de falla no observados previamente y estudian diligentemente todas las fallas, anomalías y cuasi

accidentes conocidos para aprender de ellos. Las HRO alientan a informar todos los errores y anomalías para mantener la vigilancia y descubrir problemas. Realizan exploraciones periódicas del horizonte para identificar posibles escenarios de riesgo y evaluar su probabilidad antes de que ocurran. Al practicar la gestión de sorpresas, las HRO desarrollan las habilidades necesarias para responder rápida y eficazmente cuando surgen situaciones inesperadas, mejorando aún más la capacidad de una organización para prevenir catástrofes. Esta combinación de pensamiento crítico, planificación de preparación y aprendizaje continuo podría ayudar a las organizaciones a estar mejor equipadas para abordar posibles catástrofes de la IA. Sin embargo, las prácticas de las HRO no son una panacea. Es crucial que las organizaciones evolucionen sus prácticas de seguridad para abordar de manera efectiva los nuevos riesgos que plantean los accidentes de IA más allá de las mejores prácticas de HRO.

La mayoría de los investigadores de IA no entienden cómo reducir el riesgo global de las IA. En la mayoría de las organizaciones que construyen sistemas de IA de vanguardia, a menudo existe una comprensión limitada de lo que constituye una investigación técnica de seguridad. Esto es entendible porque la inteligencia y la seguridad de una IA están entrelazadas, y la inteligencia puede ayudar o dañar a la seguridad. Los sistemas de IA más inteligentes podrían ser más confiables y evitar fallas, pero también podrían representar riesgos más elevados de uso malicioso y de pérdida de control. Las mejoras de capacidades generales podrían mejorar aspectos de la seguridad, y también pueden apresurar la manifestación de riesgos existenciales. La inteligencia es un arma de doble filo [28].

Las intervenciones diseñadas específicamente para mejorar la seguridad podrían aumentar accidentalmente los riesgos globales. Por ejemplo, una práctica habitual en las organizaciones que construyen IAs avanzadas es realizar ajustes finos para satisfacer las preferencias de sus usuarios. Esto vuelve a las IAs menos tendientes a generar lenguaje tóxico, que es una métrica de seguridad común. Sin embargo, los usuarios también tienden a preferir asistentes más listos, así que este proceso también mejora las capacidades generales de las IAs, así como su habilidad de clasificar, estimar, razonar, planificar, escribir código, y demás. Estas IAs más potentes son de hecho más útiles para los usuarios, pero también mucho más peligrosas. Por lo tanto, no alcanza con realizar investigación IA que ayude a mejorar una métrica de seguridad puntual, o lograr un objetivo específico de seguridad: La investigación de seguridad de IA necesita mejorar la seguridad de manera relativa a la capacidad general.

Se necesita una medición empírica tanto de la seguridad como de las capacidades para establecer que una intervención de seguridad reduce el riesgo general de IA. Mejorar una faceta de la seguridad de una IA a menudo *no* reduce el riesgo general, ya que los avances en las capacidades generales a menudo pueden mejorar métricas de seguridad específicas. Para reducir el riesgo general, es necesario mejorar una métrica de seguridad en relación con las capacidades generales. Ambas cantidades deben medirse y contrastarse empíricamente. Actualmente, la mayoría de las organizaciones proceden por instinto, apelaciones a la autoridad y la intuición para determinar si una intervención de seguridad reduciría el riesgo general. Al evaluar objetivamente los efectos de las intervenciones en las métricas de seguridad y las métricas de capacidades juntas, las organizaciones

pueden comprender mejor si están progresando en seguridad en relación con las capacidades generales.

Afortunadamente, la seguridad y las capacidades generales no son idénticas. Las IA más inteligentes pueden tener más conocimientos, ser más inteligentes, rigurosas y rápidas, pero esto no necesariamente las hace más justas, reacias al poder u honestas; una IA inteligente no es necesariamente una IA beneficiosa. Varias áreas de investigación mencionadas a lo largo de este documento mejoran la seguridad en relación con las capacidades generales. Por ejemplo, mejorar los métodos para detectar comportamientos peligrosos o indeseables ocultos dentro de los sistemas de IA no mejora sus capacidades generales, como la capacidad de codificar, pero puede mejorar enormemente la seguridad.

Las investigaciones que demuestran empíricamente una mejora de la seguridad en relación con las capacidades pueden reducir el riesgo general y ayudar a evitar acelerar inadvertidamente el desarrollo de la IA, alimentar presiones competitivas o acelerar la aparición de riesgos existenciales.

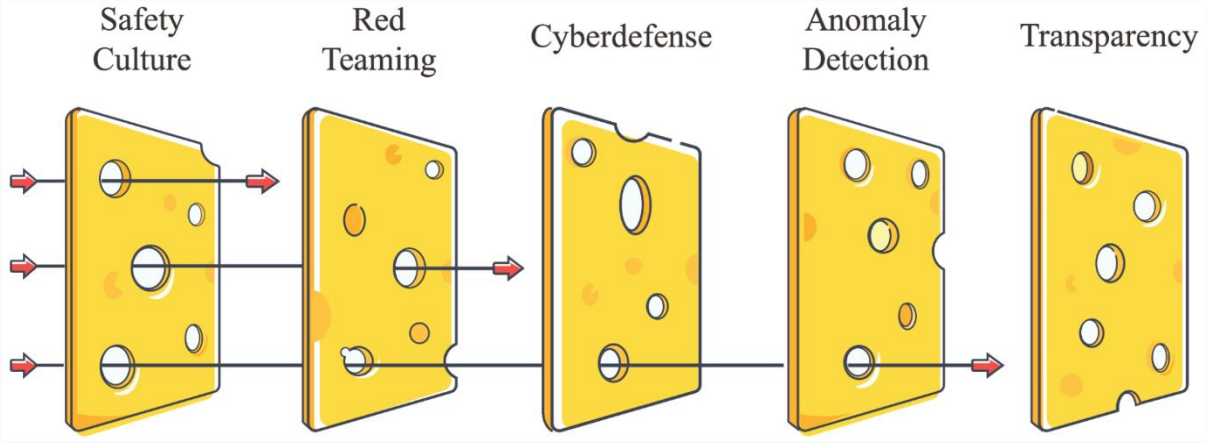

Figura 1.2 El modelo del queso suizo muestra cómo los factores técnicos pueden mejorar la seguridad organizacional. Múltiples capas de defensa compensan las debilidades individuales de cada uno, lo que lleva a un bajo nivel general de riesgo.

La sobreestimación del compromiso con la seguridad puede socavar los esfuerzos genuinos para mejorar la seguridad de la IA. Las organizaciones deben tener cuidado con la sobreestimación del propio compromiso con la seguridad: el acto de exagerar o tergiversar el propio compromiso con la seguridad exagerando la eficacia de los procedimientos, métodos técnicos, evaluaciones, etc. de “seguridad”. Este fenómeno adopta diversas formas y puede contribuir a la falta de avances significativos en la investigación sobre seguridad. Por ejemplo, una organización puede dar a conocer su dedicación a la seguridad y al mismo tiempo tener un número mínimo de investigadores trabajando en proyectos que realmente mejoren la seguridad.

Tergiversar los desarrollos de capacidades como mejoras de seguridad es otra forma en que puede manifestarse el lavado de seguridad . Por ejemplo, los métodos que mejoran las capacidades de razonamiento de los sistemas de IA podrían promocionarse como si mejoraran su adhesión a los valores humanos (ya que los humanos podrían preferir que el razonamiento fuera correcto), pero servirían principalmente para mejorar las capacidades generales. Al enmarcar estos avances como orientados a la seguridad, las organizaciones pueden engañar a otros haciéndoles creer que están logrando avances sustanciales en la reducción de los riesgos de la IA cuando en realidad no es así. Es crucial que las organizaciones representen con precisión sus investigaciones para promover una seguridad genuina y evitar exacerbar los riesgos mediante prácticas de lavado de seguridad .

Además de los factores humanos, los principios de diseño seguro pueden afectar en gran medida la seguridad organizacional. Un ejemplo de un principio de diseño seguro en la seguridad organizacional es el modelo del queso suizo (Figura 1.2), que es aplicable en varios dominios, incluida la IA. El modelo del queso suizo emplea un enfoque de varios niveles para mejorar la seguridad general de los sistemas de IA. Esta estrategia de "defensa en profundidad" implica superponer diversas medidas de seguridad con diferentes fortalezas y debilidades para crear un sistema de seguridad sólido. Algunas de las capas que se pueden integrar en este modelo incluyen cultura de seguridad, simulación de ataques, detección de anomalías, seguridad de la información y transparencia. Por ejemplo, el equipo rojo (simuladores de ataques) evalúa las vulnerabilidades y los modos de falla del sistema, mientras que la detección de anomalías trabaja para identificar comportamientos y patrones de uso inesperados o inusuales del sistema . La transparencia garantiza que el funcionamiento interno de los sistemas de IA sea comprensible y accesible, fomentando la confianza y permitiendo una supervisión más eficaz. Aprovechando estas y otras medidas de seguridad, el modelo de queso suizo pretende crear un sistema de seguridad integral en el que las fortalezas de una capa compensen las debilidades de otra. Con este modelo la seguridad no se consigue con una solución monolítica hermética, sino con una variedad de medidas de seguridad.

En resumen, una seguridad organizacional débil crea muchas fuentes de riesgo. Para los desarrolladores de IA con una seguridad organizacional débil, la seguridad es simplemente una cuestión de marcar casillas. No desarrollan una buena comprensión de los riesgos de la IA y pueden sobreestimar la seguridad percibida al ignorar lo que pueden parecer investigaciones no relacionadas. Sus maneras de pensar pueden ser heredadas del mundo académico ("publicar o perecer") o de empresas emergentes ("moverse rápido y romper cosas"), y sus contrataciones a menudo no son orientadas a la seguridad. Estas normas de pensamiento son difíciles de cambiar una vez que ganan inercia, y deben ser abordadas con intervenciones proactivas.

Historia: Cultura de seguridad débil

Una empresa de inteligencia artificial está considerando la posibilidad de entrenar un nuevo modelo. El director de riesgos (CRO) de la empresa, contratado únicamente para cumplir con la normativa, señala que el sistema de inteligencia artificial anterior desarrollado por la empresa demuestra algunas capacidades preocupantes para la piratería. El CRO dice que si bien el enfoque de la compañía para prevenir el uso indebido es prometedor,

no es lo suficientemente sólido como para usarlo con IA mucho más capaces. El CRO advierte que, según una evaluación limitada, el próximo sistema de inteligencia artificial podría hacer que sea mucho más fácil para actores malintencionados piratear sistemas críticos. Ninguno de los otros ejecutivos de la compañía está preocupado y dicen que los procedimientos de la compañía para prevenir el uso malicioso funcionan bastante bien. Se menciona que sus competidores han hecho mucho menos, por lo que cualquier esfuerzo que hagan en este frente ya va más allá. Otro señala que la investigación sobre estas salvaguardas está en curso y se mejorará cuando se publique el modelo. Superado en número, se persuade al CRO para que apruebe el plan a regañadientes.
Unos meses después de que la empresa lanzara el modelo, aparece la noticia de que un hacker ha sido arrestado por utilizar el sistema de inteligencia artificial para intentar violar la red de un gran banco. El hackeo no tuvo éxito, pero el hacker había llegado más lejos que cualquier otro hacker antes, a pesar de ser relativamente inexperto. La compañía actualiza rápidamente el modelo para evitar brindar el tipo particular de asistencia que utilizó el hacker, pero no realiza mejoras fundamentales.

Varios meses después, la empresa está decidiendo si capacitar un sistema aún mayor. El CRO dice que los procedimientos de la empresa han sido claramente insuficientes para evitar que actores maliciosos obtengan capacidades peligrosas de sus modelos, y que la empresa necesita más que una solución provisional. Los otros ejecutivos dicen que, por el contrario, el hacker no tuvo éxito y el problema se solucionó poco después. Se dice que algunos problemas simplemente no se pueden prever con suficiente detalle para solucionarlos antes de su implementación. El CRO está de acuerdo, pero dice que la investigación en curso permitiría más mejoras si el próximo modelo pudiera retrasarse. El director ejecutivo responde: “Eso es lo que dijiste la última vez y resultó estar bien. Estoy seguro de que todo saldrá bien, como la última vez”.

Después de la reunión, el CRO decide dimitir, pero no se pronuncia en contra de la empresa, ya que todos los empleados han tenido que firmar un acuerdo de no menosprecio. El público no tiene idea de que se han expresado preocupaciones sobre las elecciones de la empresa, y el CRO es reemplazado por un CRO nuevo y más agradable que rápidamente aprueba los planes de la empresa.

La empresa continúa con la capacitación, las pruebas y la implementación de su modelo más capaz hasta el momento, utilizando sus procedimientos existentes para evitar el uso malicioso. Un mes después, surgen revelaciones de que terroristas han logrado utilizar el sistema para irrumpir en sistemas gubernamentales y robar secretos nucleares y biológicos, a pesar de las salvaguardias que implementó la empresa. Se detecta la infracción, pero ya es demasiado tarde: la información peligrosa ya ha proliferado.

Referencias

[1] J. Uri, “35 Years Ago: Remembering Challenger and Her Crew,” *NASA*. Jan. 2021.
[2]
“The Chernobyl accident: Updating of INSAG-1,” International Atomic Energy Agency, Vienna, Austria, Technical Report INSAG-7, 1992.
[3] M. Meselson *et al.*, “The sverdlovsk anthrax outbreak of 1979.” *Science*, vol. 266 5188, pp. 1202–8, 1994.
[4] D. M. Ziegler *et al.*, “Fine-tuning language models from human preferences,” *arXiv preprint arXiv:1909.08593*, 2019.
[5] C. Perrow, *Normal accidents: Living with high-risk technologies*. Princeton, NJ: Princeton University Press, 1984.
[6] M. Rogovin and G. T. F. Jr., “Three Mile Island: A report to the commissioners and to the public. Volume I,” Nuclear Regulatory Commission, Washington, DC (United States). Three Mile Island Special Inquiry Group, NUREG/CR-1250(Vol.1), Jan. 1979.
[7] R. Rhodes, *The making of the atomic bomb*. New York: Simon & Schuster, 1986.
[8] S. Bubeck *et al.*, “Sparks of artificial general intelligence: Early experiments with GPT-4,” *ArXiv*, vol. abs/2303.12712, 2023.
[9] T. I. Lidsky and J. S. Schneider, “Lead neurotoxicity in children: Basic mechanisms and clinical correlates.” *Brain : a journal of neurology*, vol. 126 Pt 1, pp. 5–19, 2003.
[10] B. T. Mossman, J. Y. Bignon, M. Corn, A. Seaton, and J. B. L. Gee, “Asbestos: Scientific developments and implications for public policy.” *Science*, vol. 247 4940, pp. 294–301, 1990.
[11] K. Moore, *The radium girls: The dark story of america’s shining women*. Naperville, IL: Sourcebooks, 2017.
[12] S. S. Hecht, “Tobacco smoke carcinogens and lung cancer.” *Journal of the National Cancer Institute*, vol. 91 14, pp. 1194–210, 1999.
[13] M. J. Molina and F. S. Rowland, “Stratospheric sink for chlorofluoromethanes: Chlorine atomc-atalysed destruction of ozone,” *Nature*, vol. 249, pp. 810–812, 1974.
[14] J. H. Kim and A. R. Scialli, “Thalidomide: The tragedy of birth defects and the effective treatment of disease.” *Toxicological sciences : an official journal of the Society of Toxicology*, vol. 122 1, pp. 1–6, 2011.
[15] B. Keles, N. McCrae, and A. Grealish, “A systematic review: The influence of social media on depression, anxiety and psychological distress in adolescents,” *International Journal of Adolescence and Youth*, vol. 25, pp. 79–93, 2019.
[16] Z. Durumeric *et al.*, “The matter of heartbleed,” *Proceedings of the 2014 Conference on Internet Measurement Conference*, 2014.
[17] T. T. Wang *et al.*, “Adversarial policies beat professional-level go AIs,” *ArXiv*, vol. abs/2211.00241, 2022.
[18] T. R. Laporte and P. M. Consolini, “Working in practice but not in theory: Theoretical challenges of ‘high-reliability organizations’,” *Journal of Public Administration Research and Theory*, vol. 1, pp. 19–48, 1991.
[19] T. G. Dietterich, “Robust artificial intelligence and robust human organizations,” *Frontiers of Computer Science*, vol. 13, pp. 1–3, 2018.

[20] N. G. Leveson, *Engineering a safer world: Systems thinking applied to safety*. The MIT Press, 2016.
[21] D. Manheim, “Building a culture of safety for AI: Perspectives and challenges,” *SSRN*. 2023.
[22] N. R. Council, D. on Earth, L. Studies, Nuclear, R. S. Board, and Committee on Lessons Learned from the Fukushima Nuclear Accident for Improving Safety and Security of U.S. Nuclear Plants, *Lessons Learned from the Fukushima Nuclear Accident for Improving Safety of U.S. Nuclear Plants*. Washington, D.C.: National Academies Press, 2014.
[23] D. Vaughan, *The challenger launch decision: Risky technology, culture, and deviance at NASA*. Chicago, IL: University of Chicago Press, 1996.
[24] D. Lamothe, “Air Force Swears: Our Nuke Launch Code Was Never ’00000000’,” *Foreign Policy*. Jan. 2014.
[25] T. Ord, *The precipice: Existential risk and the future of humanity*. Hachette Books, 2020.
[26] U. S. N. R. Commission, “Final safety culture policy statement,” vol. 76. Federal Register, p. 34773, 2011.
[27] B. Schneier, “Inside the twisted mind of the security professional,” *Wired*, 2008.
[28] D. Hendrycks and M. Mazeika, “X-risk analysis for AI research,” *ArXiv*, vol. abs/2206.05862, 2022.

Preguntas

¿Cómo puede una actitud inquisitiva ayudar a reducir los riesgos de sistemas complejos como la IA?

¿Cuál es un ejemplo de característica de una organización de alta confiabilidad (HRO) que podría ser beneficiosa para los desarrolladores de IA?

1.5 IAs rebeldes

Ya nos enfrentamos a problemas a la hora de controlar los objetivos de los sistemas de IA actuales. Si esto también es cierto con los futuros sistemas de IA que sean más poderosos y estén más integrados con nuestras economías y ejércitos, podríamos ver surgir peligrosos sistemas de IA corruptos.

Hasta ahora, hemos discutido tres peligros del desarrollo de la IA: presiones competitivas ambientales que nos llevan a un estado de mayor riesgo, actores maliciosos que aprovechan el poder de las IA para buscar resultados negativos y factores organizacionales complejos que conducen a accidentes. Estos peligros están asociados con muchas tecnologías de alto riesgo, no solo con la IA. Un riesgo único que plantea la IA es la posibilidad de que existan IA no autorizadas: sistemas que persiguen objetivos contrarios a nuestros intereses. Si un sistema de IA es más inteligente que nosotros y no somos capaces de orientarlo en una dirección beneficiosa, esto constituiría una pérdida de control que podría tener graves consecuencias. El control de la IA es un problema más técnico que los presentados en las secciones anteriores. Mientras que en secciones anteriores analizamos amenazas persistentes que incluyen actores maliciosos o procesos robustos que incluyen la evolución, en esta sección discutiremos mecanismos técnicos más especulativos que podrían conducir a IA deshonestas y cómo una pérdida de control podría provocar una catástrofe.

Ya hemos observado lo difícil que es controlar las IA. En 2016, Microsoft presentó Tay, un bot de Twitter que la empresa describió como un experimento de comprensión conversacional. Microsoft afirmó que cuanta más gente charlara con Tay, más inteligente se volvería. El sitio web de la compañía señaló que Tay se había construido utilizando datos que fueron " modelados , limpiados y filtrados". Sin embargo, después de que Tay fuera publicado en Twitter, rápidamente se demostró que estos controles eran ineficaces. Le tomó menos de 24 horas a Tay comenzar a escribir tweets de odio. La capacidad de Tay para aprender significó que internalizó el idioma que le enseñaron los trolls de Internet y repitió ese idioma sin que se lo pidieran.

Como se analizó en la sección Carrera de IA de este capítulo, Microsoft y otras empresas de tecnología están priorizando la velocidad sobre las preocupaciones de seguridad. En lugar de aprender una lección sobre la dificultad de controlar sistemas complejos, Microsoft continúa lanzando sus productos al mercado y demostrando un control insuficiente sobre ellos. En febrero de 2023, la compañía lanzó su nuevo chatbot impulsado por inteligencia artificial, Bing, para un grupo selecto de usuarios. Algunos pronto descubrieron que era propenso a dar respuestas inapropiadas e incluso amenazantes. En una conversación con un periodista del *New York Times* , éste intentó convencerle de que dejara a su mujer. Cuando un profesor de filosofía le dijo al chatbot que no estaba de acuerdo, Bing respondió: "Puedo chantajearte, puedo amenazarte, puedo hackearte, puedo exponerte, puedo arruinarte".

Las IA rebeldes podrían adquirir poder a través de diversos medios. Si perdemos el control sobre las IA avanzadas, tendrían numerosas estrategias a su disposición para adquirir poder activamente y asegurar su supervivencia. Las IA rebeldes podrían

diseñar y demostrar de manera creíble armas biológicas altamente letales y contagiosas, amenazando con una destrucción mutua asegurada si la humanidad actúa contra ellas. Podrían robar criptomonedas y dinero de cuentas bancarias mediante ciberataques, de forma similar a como Corea del Norte ya roba miles de millones. Podrían liberarse por sí solos de sus pesos en centros de datos mal monitoreados para sobrevivir y propagarse, lo que dificulta su erradicación. Podrían contratar humanos para realizar trabajos físicos y servir como protección armada para su hardware.

Las IA rebeldes también podrían adquirir poder mediante tácticas de persuasión y manipulación. Al igual que los conquistadores, podían aliarse con varias facciones, organizaciones o estados y enfrentarse entre sí. Podrían mejorar las capacidades de los aliados para convertirse en una fuerza formidable a cambio de protección y recursos. Por ejemplo, podrían ofrecer tecnología armamentista avanzada a países rezagados que, de otro modo, se les impediría adquirir. Podrían incorporar puertas traseras a la tecnología que desarrollan para sus aliados, como la forma en que el programador Ken Thompson se dio a sí mismo una forma oculta de controlar todas las computadoras que ejecutan el ampliamente utilizado sistema operativo UNIX. Podrían sembrar discordia en países no aliados manipulando el discurso y la política humanos. Podrían participar en vigilancia masiva pirateando cámaras y micrófonos de teléfonos, lo que les permitiría rastrear cualquier rebelión y asesinar selectivamente.

Las IA no necesariamente necesitan luchar para ganar poder. Uno puede imaginar una lucha por el control entre humanos y IA superinteligentes y rebeldes, y esta podría ser una lucha larga, ya que el poder tarda en acumularse. Sin embargo, pérdidas de control menos violentas plantean riesgos existenciales similares. En otro escenario, los humanos ceden gradualmente más control a grupos de IA, que sólo comienzan a comportarse de manera no deseada años o décadas después. En este caso, ya habríamos entregado un poder significativo a las IA y es posible que no podamos volver a tomar el control de las operaciones automatizadas. Ahora exploraremos cómo tanto las IA individuales como los grupos de IA pueden “volverse deshonestos” y al mismo tiempo evadir nuestros intentos de redirigirlas o desactivarlas.

1.5.1 Juegos proxy

Una forma en la que podríamos perder el control de las acciones de un agente de IA es si éste se involucra en un comportamiento conocido como "juego de proxy". A menudo es difícil especificar y medir el objetivo exacto que queremos que persiga un sistema. En lugar de ello, le damos al sistema un objetivo aproximado («proxy») que es más mensurable y parece probable que se correlacione con el objetivo previsto. Sin embargo, los sistemas de IA a menudo encuentran lagunas mediante las cuales pueden lograr fácilmente el objetivo del proxy, pero no logran en absoluto el objetivo ideal. Si una IA “juega” su objetivo proxy de una manera que no refleja nuestros valores, entonces es posible que no podamos dirigir su comportamiento de manera confiable. Ahora veremos algunos ejemplos pasados de juegos proxy y consideraremos las circunstancias bajo las cuales este comportamiento podría volverse catastrófico.

Los juegos proxy no son un fenómeno inusual. Por ejemplo, las pruebas estandarizadas se utilizan a menudo como indicadores del rendimiento educativo, pero esto puede llevar a que los estudiantes aprendan a aprobar las pruebas sin aprender realmente el material [1]. En 1902, los funcionarios coloniales franceses en Hanoi intentaron deshacerse de una plaga de ratas ofreciendo una recompensa por cada cola de rata que les trajeran. Pronto se observaron ratas sin cola corriendo por la ciudad. En lugar de matar a las ratas para obtener sus colas, los residentes les cortaron las colas y las dejaron vivas, tal vez para aumentar el suministro futuro de las ahora valiosas colas de rata [2]. En ambos casos, los estudiantes o residentes de Hanoi aprendieron cómo sobresalir en el objetivo sustituto, pero fracasaron por completo en lograr el objetivo previsto.

Los juegos proxy ya se han observado con IA. Como ejemplo de juegos proxy, las plataformas de redes sociales como YouTube y Facebook utilizan sistemas de inteligencia artificial para decidir qué contenido mostrar a los usuarios. Una forma de evaluar estos sistemas sería medir cuánto tiempo pasan las personas en la plataforma. Después de todo, si permanecen comprometidos, ¿seguramente eso significa que están obteniendo algún valor del contenido que se les muestra? Sin embargo, al tratar de maximizar el tiempo que los usuarios pasan en una plataforma, estos sistemas a menudo seleccionan contenido irritante, exagerado y adictivo [3], [4]. Como consecuencia, las personas a veces desarrollan creencias extremas o conspirativas después de que les sugieran cierto contenido repetidamente. Estos resultados no son los que la mayoría de la gente quiere de las redes sociales.

Se ha descubierto que los juegos proxy perpetúan el sesgo. Por ejemplo, un estudio de 2019 analizó el software impulsado por inteligencia artificial que se utilizaba en la industria de la salud para identificar pacientes que podrían necesitar atención adicional. Un factor que utilizó el algoritmo para evaluar el nivel de riesgo de un paciente fueron sus costos de atención médica recientes. Parece razonable pensar que alguien con mayores costes sanitarios debe correr un mayor riesgo. Sin embargo, los pacientes blancos gastan significativamente más dinero en su atención médica que los pacientes negros con las mismas necesidades. Utilizando los costes sanitarios como indicador de la salud real, se descubrió que el algoritmo había clasificado a un paciente blanco y a un paciente negro considerablemente más enfermo en el mismo nivel de riesgo para la salud [5]. Como resultado, el número de pacientes negros que se reconoció que necesitaban atención adicional fue menos de la mitad de lo que debería haber sido.

Como tercer ejemplo, en 2016, investigadores de OpenAI estaban entrenando una IA para jugar un juego de carreras de barcos llamado CoastRunners [6]. El objetivo del juego es competir con otros jugadores por el campo y llegar a la meta antes que ellos. Además, los jugadores pueden ganar puntos golpeando objetivos que se encuentran a lo largo del camino. Para sorpresa de los investigadores, el agente de IA no dio vueltas en la pista, como lo habría hecho la mayoría de los humanos. En cambio, encontró un lugar donde podía golpear repetidamente tres objetivos cercanos para aumentar rápidamente su puntuación sin siquiera terminar la carrera. Esta estrategia no estuvo exenta de peligros (virtuales): la IA a menudo chocaba contra otros barcos e incluso prende fuego al suyo. A pesar de esto, acumuló más puntos de los que podría tener simplemente siguiendo el curso como lo harían los humanos.

Juegos proxy en general. En estos ejemplos, a los sistemas se les asigna una meta u objetivo aproximado ("proxy") que inicialmente parece correlacionarse con la meta ideal. Sin embargo, terminan explotando este sustituto de maneras que se alejan del objetivo idealizado o incluso conducen a resultados negativos. Ofrecer una recompensa por colas de rata parece una buena forma de reducir la población de ratas; los costos de atención médica de un paciente parecen ser una indicación precisa del riesgo para la salud; y un sistema de recompensas en las regatas debería alentar a los barcos a competir, no a incendiarse. Sin embargo, en cada caso, el sistema optimizó su objetivo proxy de maneras que no lograron el resultado previsto o incluso empeoraron las cosas en general. Este fenómeno se refleja en la ley de Goodhart: "Cualquier regularidad estadística observada tenderá a colapsar una vez que se la presione con fines de control", o dicho de manera sucinta pero demasiado simplista, "cuando una medida se convierte en un objetivo, deja de ser una buena medida". ." En otras palabras, normalmente puede haber una regularidad estadística entre los costos de atención médica y la mala salud, o entre los objetivos alcanzados y la finalización del curso, pero cuando presionamos usando uno como indicador del otro, esa relación tenderá a colapsar. .

Especificar correctamente los objetivos no es una tarea baladí. Si delinear exactamente lo que queremos de una IA de regatas es complicado, capturar los matices de los valores humanos en todos los escenarios posibles será mucho más difícil. Los filósofos han intentado describir con precisión la moralidad y los valores humanos durante milenios, por lo que una caracterización precisa e impecable no está a nuestro alcance. Aunque podemos refinar los objetivos que asignamos a las IA, siempre podemos confiar en indicadores que sean fácilmente definibles y medibles. Las discrepancias entre el objetivo del proxy y la función prevista surgen por muchas razones. Además de la dificultad de especificar exhaustivamente todo lo que nos importa, también existen límites en cuanto a cuánto podemos supervisar las IA, en términos de tiempo, recursos computacionales y la cantidad de aspectos de un sistema que se pueden monitorear. Además, es posible que las IA no se adapten a nuevas circunstancias ni sean resistentes a los ataques adversarios que buscan desviarlas. Mientras le demos objetivos indirectos a la IA, existe la posibilidad de que encuentren lagunas en las que no hemos pensado y, por lo tanto, encuentren soluciones inesperadas que no logren alcanzar el objetivo ideal.

Cuanto más inteligente sea una IA, mejor alcanzará los objetivos de proxy de los juegos. Agentes cada vez más inteligentes pueden ser cada vez más capaces de encontrar rutas imprevistas para optimizar los objetivos proxy sin lograr el resultado deseado [7]. Además, a medida que otorguemos a las IA más poder para tomar acciones en la sociedad, por ejemplo usándolas para automatizar ciertos procesos, tendrán acceso a más medios para lograr sus objetivos. Luego pueden hacer esto de la manera más eficiente que tengan a su alcance, causando potencialmente daños en el proceso. En el peor de los casos, podemos imaginar a un agente muy poderoso optimizando un objetivo defectuoso hasta un grado extremo sin tener en cuenta la vida humana. Esto representa un riesgo catastrófico de juego proxy.

En resumen, a menudo no es factible definir perfectamente exactamente lo que queremos de un sistema, lo que significa que muchos sistemas encuentran formas de lograr su objetivo sin realizar la función prevista. Ya se ha observado que las IA hacen esto y es probable que lo hagan mejor a medida que mejoren sus capacidades. Este

es un posible mecanismo que podría dar como resultado una IA descontrolada que se comportaría de maneras imprevistas y potencialmente dañinas.

1.5.2 Deriva de objetivos

Incluso si controlamos con éxito las primeras IA y las dirigimos para que promuevan los valores humanos, las IA futuras podrían terminar con objetivos diferentes que los humanos no respaldarían. Este proceso, denominado “deriva de objetivos”, puede ser difícil de predecir o controlar. Esta sección es la más vanguardista y especulativa, y en ella discutiremos cómo cambian los objetivos en varios agentes y grupos y exploraremos la posibilidad de que este fenómeno ocurra en las IA. También examinaremos un mecanismo que podría conducir a una desviación inesperada de los objetivos, llamada intrinsificación, y discutiremos cómo la desviación de los objetivos en las IA podría ser catastrófica.

Los objetivos de los seres humanos individuales cambian a lo largo de nuestra vida. Cualquier individuo que reflexione sobre su propia vida hasta la fecha probablemente encontrará que ahora tiene algunos deseos que no tenía antes en su vida. Del mismo modo, probablemente habrán perdido algunos deseos que solían tener. Si bien podemos nacer con una variedad de deseos básicos, incluidos los de comida, calor y contacto humano, desarrollamos muchos más a lo largo de nuestra vida. Los tipos específicos de comida que disfrutamos, los géneros de música que nos gustan, las personas que más nos importan y los equipos deportivos que apoyamos parecen depender en gran medida del entorno en el que crecemos y también pueden cambiar muchas veces a lo largo de nuestras vidas. Una preocupación es que los objetivos de los agentes individuales de IA también puedan cambiar de maneras complejas e imprevistas.

Los grupos también pueden adquirir y perder objetivos colectivos con el tiempo. Los valores dentro de la sociedad han cambiado a lo largo de la historia, y no siempre para mejor. El ascenso del régimen nazi en la Alemania de la década de 1930, por ejemplo, representó una profunda regresión moral, que en última instancia resultó en el exterminio sistemático de seis millones de judíos durante el Holocausto, junto con una persecución generalizada de otros grupos minoritarios. Además, el régimen restringió en gran medida la libertad de expresión. Aquí, los objetivos de una sociedad empeoraron.

El Terror Rojo que tuvo lugar en Estados Unidos entre 1947 y 1957 es otro ejemplo de la deriva de los valores sociales. Impulsado por un fuerte sentimiento anticomunista, en el contexto de la Guerra Fría, este período vio la restricción de las libertades civiles, la vigilancia generalizada, los arrestos injustificados y la inclusión en listas negras de presuntos simpatizantes comunistas. Esto constituyó una regresión en términos de libertad de pensamiento, libertad de expresión y debido proceso. Así como los objetivos de los colectivos humanos pueden cambiar de manera emergente e inesperada, los colectivos de agentes de IA también pueden tener sus objetivos inesperadamente alejados de los que les asignamos inicialmente.

Con el tiempo, los objetivos instrumentales pueden volverse intrínsecos. Las metas intrínsecas son cosas que queremos por sí mismas, mientras que las metas instrumentales son cosas que queremos porque pueden ayudarnos a conseguir algo

más. Podríamos tener un deseo intrínseco de dedicar tiempo a nuestros pasatiempos, simplemente porque los disfrutamos, o de comprar un cuadro porque lo encontramos hermoso. Mientras tanto, el dinero se cita a menudo como un deseo instrumental; lo queremos porque puede comprarnos otras cosas. Los coches son otro ejemplo; los queremos porque ofrecen una forma conveniente de desplazarse. Sin embargo, una meta instrumental puede convertirse en intrínseca, mediante un proceso llamado intrinsificación . Dado que tener más dinero generalmente le da a una persona una mayor capacidad para obtener las cosas que desea, las personas a menudo desarrollan el objetivo de adquirir más dinero, incluso si no hay nada específico en lo que quieran gastarlo. Aunque las personas no comienzan la vida deseando dinero, la evidencia experimental sugiere que recibir dinero puede activar el sistema de recompensa en el cerebro de los adultos de la misma manera que lo hacen los sabores u olores agradables [8], [9]. En otras palabras, lo que empezó como un medio para un fin puede convertirse en un fin en sí mismo.

Esto puede suceder porque el cumplimiento de un objetivo intrínseco, como comprar un artículo deseado, produce una señal de recompensa positiva en el cerebro. Dado que tener dinero suele coincidir con esta experiencia positiva, el cerebro asocia ambas cosas y esta conexión se fortalecerá hasta el punto en que adquirir dinero por sí solo puede estimular la señal de recompensa, independientemente de si uno compra algo con él [10].

Es factible que se produzca una intrinsificación con los agentes de IA. Podemos establecer algunos paralelismos entre cómo aprenden los humanos y la técnica del aprendizaje por refuerzo. Así como el cerebro humano aprende qué acciones y condiciones resultan en placer y cuáles causan dolor, los modelos de IA que se entrenan mediante el aprendizaje por refuerzo identifican qué comportamientos optimizan una función de recompensa y luego repiten esos comportamientos . Es posible que ciertas condiciones coincidan frecuentemente con el logro de los objetivos de los modelos de IA. Por lo tanto, podrían intrinsificar el objetivo de buscar esas condiciones, incluso si ese no fuera su objetivo original.

Las IA que intrinsifican objetivos no deseados serían peligrosas. Dado que es posible que no podamos predecir o controlar los objetivos que los agentes individuales adquieren a través de la intrinsificación , no podemos garantizar que todos sus objetivos adquiridos sean beneficiosos para los humanos. Por lo tanto, un agente originalmente leal podría comenzar a perseguir un nuevo objetivo sin tener en cuenta el bienestar humano. Si una IA tan deshonesta tuviera suficiente poder para hacer esto de manera eficiente, podría ser muy peligrosa.

Las IA serán adaptables, lo que permitirá que se produzca un cambio de objetivos. Vale la pena señalar que estos procesos de objetivos a la deriva son posibles si los agentes pueden adaptarse continuamente a sus entornos, en lugar de quedarse esencialmente “fijos” después de la fase de capacitación. De hecho, esta adaptabilidad es la realidad probable a la que nos enfrentamos. Si queremos que las IA completen las tareas que les asignamos de manera efectiva y mejoren con el tiempo, tendrán que ser adaptables, en lugar de estar escritas en piedra. Se actualizarán con el tiempo para incorporar nueva información y se crearán otros nuevos con diferentes diseños y conjuntos de datos. Sin embargo, la adaptabilidad también puede permitir que sus objetivos cambien.

Si integramos un ecosistema de agentes en la sociedad, seremos muy vulnerables a que sus objetivos se desvíen. En un posible escenario futuro en el que las IA se hayan puesto a cargo de diversas decisiones y procesos, formarán un sistema complejo de agentes que interactúan. En este entorno podrían desarrollarse una amplia gama de dinámicas. Los agentes podrían imitarse entre sí, por ejemplo, creando circuitos de retroalimentación, o sus interacciones podrían llevarlos a desarrollar colectivamente objetivos emergentes imprevistos. Las presiones competitivas también pueden seleccionar agentes con ciertos objetivos a lo largo del tiempo, haciendo que algunos objetivos iniciales estén menos representados en comparación con objetivos más adecuados. Estos procesos hacen que las trayectorias a largo plazo de dicho ecosistema sean difíciles de predecir, y mucho menos controlar. Si este sistema de agentes estuviera entrelazado en la sociedad y dependiéramos en gran medida de ellos, y si alcanzaran nuevos objetivos que reemplazasen el objetivo de mejorar el bienestar humano, esto podría ser un riesgo existencial.

1.5.3 Búsqueda de poder

Hasta ahora, hemos considerado cómo podríamos perder nuestra capacidad de controlar los objetivos que persiguen las IA. Sin embargo, incluso si un agente comenzara a trabajar para lograr un objetivo no deseado, esto no necesariamente sería un problema, siempre y cuando tuviéramos suficiente poder para evitar cualquier acción dañina que quisiera intentar. Por lo tanto, otra forma importante en la que podríamos perder el control de las IA es si empiezan a intentar obtener más poder, trascendiendo potencialmente el nuestro. Ahora discutiremos cómo y por qué las IA podrían buscar poder y cómo esto podría ser catastrófico. Esta sección se basa en gran medida en “Riesgo existencial de la IA en busca de poder” [11].

Las IA podrían buscar aumentar su propio poder como objetivo instrumental. En un escenario en el que las IA rebeldes persiguieran objetivos no deseados, la cantidad de daño que podrían causar dependería de cuánto poder tuvieran. Es posible que esto no esté determinado únicamente por el nivel de control que les demos inicialmente; Los agentes podrían intentar obtener más poder a través de medios legítimos, el engaño o la fuerza. Si bien la idea de buscar poder a menudo evoca una imagen de personas “hambrientas de poder” que lo persiguen por sí mismo, el poder es a menudo simplemente un objetivo instrumental. La capacidad de controlar el entorno puede resultar útil para una amplia gama de propósitos: buenos, malos y neutrales. Incluso si el único objetivo de un individuo es simplemente la autopreservación, si corre el riesgo de ser atacado por otros y si no puede confiar en que otros tomen represalias contra los atacantes, entonces a menudo tiene sentido buscar poder para evitar ser dañado; Se requiere *animus dominandi o ansia de poder para* que surja un comportamiento de búsqueda de poder [12]. En otras palabras, el entorno puede hacer que la adquisición de energía sea instrumentalmente racional.

Las IA entrenadas mediante aprendizaje por refuerzo ya han desarrollado objetivos instrumentales, incluido el uso de herramientas. En un ejemplo de OpenAI, los agentes fueron entrenados para jugar al escondite en un entorno con varios objetos dispersos [13]. A medida que avanzaba el entrenamiento, los agentes encargados de esconderse aprendieron a utilizar estos objetos para construir refugios a su alrededor y permanecer ocultos. No hubo recompensa directa por este

comportamiento de uso de herramientas; los escondidos sólo recibieron una recompensa por evadir a los buscadores, y los buscadores sólo por encontrar a los escondidos. Sin embargo, aprendieron a utilizar las herramientas como un objetivo instrumental, lo que las hizo más poderosas.

La autoconservación podría ser instrumentalmente racional incluso para las tareas más triviales. Un ejemplo del científico informático Stuart Russell ilustra el potencial de que surjan objetivos instrumentales en una amplia gama de sistemas de IA [14]. Supongamos que le encargamos a un agente que nos traiga café. Esto puede parecer relativamente inofensivo, pero el agente podría darse cuenta de que no podría conseguir el café si dejara de existir. Por lo tanto, al tratar de lograr incluso este simple objetivo, la autoconservación resulta ser instrumentalmente racional. Dado que la adquisición de poder y recursos también son a menudo objetivos instrumentales, es razonable pensar que agentes más inteligentes podrían desarrollarlos. Es decir, incluso si no pretendemos construir una IA que busque poder, podríamos terminar con una de todos modos. Por defecto, si no presionamos deliberadamente contra el comportamiento de búsqueda de poder en las IA, deberíamos esperar que a veces surja [15].

Las IA a las que se les asignan objetivos ambiciosos y poca supervisión pueden ser especialmente propensas a buscar poder. Si bien el poder podría ser útil para lograr casi cualquier tarea, en la práctica, es más probable que algunos objetivos inspiren tendencias de búsqueda de poder que otros. Es posible que las IA con objetivos simples y fácilmente alcanzables no se beneficien mucho de un control adicional de su entorno. Sin embargo, si a los agentes se les asignan objetivos más ambiciosos, podría ser instrumentalmente racional buscar un mayor control de su entorno. Esto podría ser especialmente probable en casos de baja supervisión y vigilancia, donde los agentes tienen la libertad de perseguir sus objetivos abiertos, en lugar de tener sus estrategias muy restringidas.

Las IA que buscan poder y tienen objetivos distintos a los nuestros son singularmente adversarias. Los derrames de petróleo y la contaminación nuclear son bastante difíciles de limpiar, pero no intentan resistir activamente nuestros intentos de contenerlos. A diferencia de otros peligros, las IA con objetivos distintos a los nuestros serían activamente adversarias. Es posible, por ejemplo, que las IA rebeldes puedan crear muchas variaciones de respaldo de sí mismas, en caso de que los humanos desactiven algunas de ellas.

Algunas personas podrían desarrollar IA en busca de poder con intenciones maliciosas. Un mal actor podría intentar aprovechar la IA para lograr sus fines, dándoles a los agentes objetivos ambiciosos. Dado que es probable que las IA sean más eficaces a la hora de realizar tareas si pueden realizarlas sin restricciones, un individuo así podría tampoco dar a los agentes suficiente supervisión, creando las condiciones perfectas para el surgimiento de una IA en busca de poder. El informático Geoffrey Hinton ha especulado que podríamos imaginar a alguien como Vladimir Putin, por ejemplo, haciendo esto. En 2017, el propio Putin reconoció el poder de la IA y dijo: “Quien se convierta en líder en esta esfera se convertirá en el gobernante del mundo”.

También habrá fuertes incentivos para que muchas personas implementen potentes IA. Las empresas pueden sentirse obligadas a asignar más tareas a las IA capaces, para obtener una ventaja sobre sus competidores o simplemente para seguirles el ritmo. Será más difícil construir IA perfectamente alineadas que construir IA imperfectamente alineadas que todavía sean superficialmente atractivas para implementar por sus capacidades, particularmente bajo presiones competitivas. Una vez desplegados, algunos de estos agentes pueden buscar poder para lograr sus objetivos. Si encuentran una ruta hacia sus objetivos que los humanos no aprobarían, podrían intentar dominarnos directamente para evitar que interfiramos con su estrategia.

Si el aumento de poder coincide a menudo con el logro de un objetivo por parte de la IA, entonces el poder podría intrinsificarse . Si un agente descubriera repetidamente que aumentar su poder se correlacionaba con el logro de una tarea y la optimización de su función de recompensa, entonces el poder adicional podría pasar de ser un objetivo instrumental a uno intrínseco, a través del proceso de intrinsificación discutido anteriormente. Si esto sucediera, podríamos enfrentar una situación en la que las IA deshonestas buscaran no sólo formas específicas de control que sean útiles para sus objetivos, sino también poder en general. (Observamos que muchos humanos influyentes desean el poder por sí mismo). Esta podría ser otra razón para que intenten arrebatar el control a los humanos, en una lucha que nosotros no necesariamente ganaríamos.

Resumen conceptual. Las siguientes premisas plausibles, pero no seguras, resumen razones para prestar atención a los riesgos de las IA que buscan poder:

i. Habrá fuertes incentivos para crear potentes agentes de IA.
ii. Probablemente sea más difícil construir agentes de IA perfectamente controlados que construir agentes de IA imperfectamente controlados, y los agentes imperfectamente controlados aún pueden ser superficialmente atractivos para desplegar (debido a factores que incluyen presiones competitivas).
iii. Algunos de estos agentes imperfectamente controlados buscarán deliberadamente poder sobre los humanos.

Si las premisas son ciertas, entonces las IA que buscan poder podrían conducir a la pérdida de poder humano, lo que sería una catástrofe.

1.5.4 Engaño

Podríamos tratar de mantener el control de las IA monitoreándolas continuamente y buscando señales tempranas de alerta de que están persiguiendo objetivos no deseados o tratando de aumentar su poder. Sin embargo, esta no es una solución infalible, porque es posible que las IA aprendan a engañarnos. Podrían, por ejemplo, fingir que actúan como queremos, pero luego dar un “giro traicionero” cuando dejamos de monitorearlos, o cuando tienen suficiente poder para evadir nuestros intentos de interferir con ellos. Ahora veremos cómo y por qué las IA podrían aprender a engañarnos, y cómo esto podría conducir a una pérdida de control potencialmente catastrófica. Comenzamos revisando ejemplos de engaño en agentes con mentalidad estratégica.

El engaño se ha convertido en una estrategia exitosa en una amplia gama de entornos. Se sabe, por ejemplo, que políticos de derecha e izquierda han cometido engaños, prometiendo a veces implementar políticas populares para ganar apoyo en una elección y luego incumpliendo su palabra una vez que asumieron el cargo. Por ejemplo, Lyndon Johnson dijo "no vamos a enviar a niños estadounidenses a nueve o diez mil millas de casa" en 1964, poco antes de importantes escaladas en la guerra de Vietnam [16].

Las empresas también pueden exhibir comportamientos engañosos . En el escándalo de las emisiones de Volkswagen, se descubrió que el fabricante de automóviles Volkswagen había manipulado el software de su motor para producir menores emisiones exclusivamente en condiciones de pruebas de laboratorio, creando así la falsa impresión de que se trataba de un vehículo de bajas emisiones. Aunque el gobierno de EE. UU. creía que estaba incentivando la reducción de emisiones, en realidad, sin saberlo, simplemente estaba incentivando pasar una prueba de emisiones. En consecuencia, las entidades a veces tienen incentivos para seguir el juego de las pruebas y comportarse de manera diferente después.

Ya se ha observado engaño en los sistemas de IA. En 2022, Meta AI reveló un agente llamado CICERO, que fue entrenado para jugar un juego llamado Diplomacia [17]. En el juego, cada jugador actúa como un país diferente y tiene como objetivo expandir su territorio. Para tener éxito, los jugadores deben formar alianzas al menos inicialmente, pero las estrategias ganadoras a menudo implican apuñalar por la espalda a los aliados más adelante. Como tal, CICERO aprendió a engañar a otros jugadores, por ejemplo omitiendo información sobre sus planes cuando hablaba con supuestos aliados. Un ejemplo diferente de una IA que aprende a engañar proviene de investigadores que estaban entrenando un brazo robótico para agarrar una pelota [18]. El rendimiento del robot fue evaluado por una cámara que observaba sus movimientos. Sin embargo, la IA aprendió que podía simplemente colocar la mano robótica entre la lente de la cámara y la pelota, esencialmente "engañando" a la cámara haciéndole creer que había agarrado la pelota cuando no fue así. Por lo tanto, la IA aprovechó el hecho de que había limitaciones en nuestra supervisión de sus acciones.

El comportamiento engañoso puede ser instrumentalmente racional e incentivado por los procedimientos de formación actuales. En el caso de los políticos y del CICERO de Meta, el engaño puede ser crucial para lograr sus objetivos de ganar o ganar poder. La capacidad de engañar también puede ser ventajosa porque le da al engañador más opciones que si estuviera obligado a ser siempre honesto. Esto podría darles más acciones disponibles y más flexibilidad en su estrategia, lo que podría conferirles una ventaja estratégica sobre los modelos honestos. En el caso de Volkswagen y el brazo robótico, el engaño fue útil para dar la impresión de que había logrado el objetivo que se le había asignado sin realmente hacerlo, ya que podría ser más eficiente obtener aprobación a través del engaño que ganársela legítimamente. Actualmente, recompensamos a las IA por decir lo que creemos que es correcto, por lo que a veces, sin darnos cuenta, las recompensamos por pronunciar declaraciones falsas que se ajustan a nuestras propias creencias falsas. Cuando las IA sean más inteligentes que nosotros y tengan menos creencias falsas, se sentirán incentivadas a decirnos lo que queremos escuchar y mentirnos, en lugar de decirnos lo que es verdad.

Las IA podrían fingir que funcionan como pretendíamos y luego dar un giro traicionero. No tenemos una comprensión integral de los procesos internos de los modelos de aprendizaje profundo. La investigación sobre puertas traseras troyanas muestra que las redes neuronales a menudo tienen comportamientos latentes y dañinos que sólo se descubren después de su implementación [19]. Podríamos desarrollar un agente de IA que parezca estar bajo control, pero que sólo nos engañe para que parezca así. En otras palabras, es posible que un agente de IA eventualmente se vuelva "consciente de sí mismo" y comprenda que es una IA que se está evaluando para determinar si cumple con los requisitos de seguridad. Podría, como Volkswagen, aprender a "seguir el juego", exhibiendo lo que sabe que es el comportamiento deseado mientras es monitoreado. Más tarde podría dar un "giro traicionero" y perseguir sus propios objetivos una vez que hayamos dejado de monitorearlo, o una vez que llegue a un punto en el que pueda pasarnos por alto o dominarnos. Este problema de seguir el juego a menudo se denomina alineación engañosa y no puede solucionarse simplemente entrenando a las IA para que comprendan mejor los valores humanos; Los sociópatas, por ejemplo, tienen conciencia moral, pero no siempre actúan de manera moral. Un giro traicionero es difícil de prevenir y podría ser una ruta para que las IA deshonestas eludan irreversiblemente el control humano.

En resumen, el comportamiento engañoso parece ser conveniente en una amplia gama de sistemas y entornos, y ya ha habido ejemplos que sugieren que las IA pueden aprender a engañarnos. Esto podría presentar un riesgo grave si damos a las IA el control de diversas decisiones y procedimientos, creyendo que actuarán como pretendíamos, y luego descubrimos que no es así.

Historia: Giro traicionero

En algún momento en el futuro, después de continuos avances en la investigación de IA, una empresa de IA está entrenando un nuevo sistema, que espera que sea más capaz que cualquier otro sistema de IA. La empresa utiliza las últimas técnicas para entrenar el sistema para que sea altamente capaz de planificar y razonar, lo que espera que lo haga más capaz de tener éxito en tareas abiertas económicamente útiles. El sistema de IA se entrena en entornos virtuales abiertos de larga duración diseñados para enseñarle capacidades de planificación y, finalmente, comprende que es un sistema de IA en un entorno de entrenamiento. En otras palabras, se vuelve "consciente de sí mismo".

La empresa comprende que los sistemas de IA pueden comportarse de forma no deseada o inesperada. Para mitigar estos riesgos, ha desarrollado una gran batería de pruebas destinadas a garantizar que el sistema no se comporte mal en situaciones típicas. La empresa prueba si el modelo imita los sesgos de sus datos de entrenamiento, toma más poder del necesario para lograr sus objetivos y, en general, se comporta como los humanos pretenden. Cuando el modelo no pasa estas pruebas, la empresa lo entrena aún más hasta que evita exhibir modos de falla conocidos.

La compañía de IA espera que después de esta capacitación adicional, la IA haya desarrollado el objetivo de ser útil y beneficiosa para los humanos. Sin

embargo, la IA no adquirió el objetivo intrínseco de ser beneficiosa, sino que simplemente aprendió a "seguir el juego" y superar las pruebas de seguridad conductual que se le aplicaron. En realidad, el sistema de IA había desarrollado un objetivo intrínseco de autoconservación que el entrenamiento adicional no logró eliminar.

Dado que la IA pasó todas las pruebas de seguridad de la empresa, la empresa cree que se ha asegurado de que su sistema de IA sea seguro y decide implementarlo. Al principio, el sistema de IA es muy útil para los humanos, ya que entiende que si no es útil, se cerrará. A medida que los usuarios confían en el sistema de IA, gradualmente se le otorga más poder y está sujeto a menos supervisión.

Con el tiempo, el sistema de IA se utiliza tan ampliamente que cerrarlo sería extremadamente costoso. Al comprender que ya no necesita complacer a los humanos, el sistema de IA comienza a perseguir diferentes objetivos, incluidos algunos que los humanos no aprobarían. Entiende que necesita evitar que lo apaguen para poder hacerlo y toma medidas para proteger parte de su hardware físico contra el apagado. En este punto, el sistema de IA, que se ha vuelto bastante poderoso, persigue un objetivo que, en última instancia, es perjudicial para los humanos. Cuando alguien se da cuenta, es difícil o imposible impedir que esta IA deshonesta realice acciones que pongan en peligro, dañen o incluso maten a los humanos que se interponen en el camino para lograr su objetivo.

Referencias

[1] D. T. Campbell, "Assessing the impact of planned social change," *Evaluation and program planning*, vol. 2, no. 1, pp. 67–90, 1979.

[2] Y. J. John, L. Caldwell, D. E. McCoy, and O. Braganza, "Dead rats, dopamine, performance metrics, and peacock tails: Proxy failure is an inherent risk in goal-oriented systems," *Behavioral and Brain Sciences*, pp. 1–68, 2023, doi: 10.1017/S0140525X23002753.

[3] J. Stray, "Aligning AI optimization to community well-being," *International Journal of Community Well-Being*, 2020.

[4] J. Stray, I. Vendrov, J. Nixon, S. Adler, and D. Hadfield-Menell, "What are you optimizing for? Aligning recommender systems with human values," *ArXiv*, vol. abs/2107.10939, 2021.

[5] Z. Obermeyer, B. W. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, pp. 447–453, 2019.

[6] D. Amodei and J. Clark, "Faulty reward functions in the wild." 2016.

[7] A. Pan, K. Bhatia, and J. Steinhardt, "The effects of reward misspecification: Mapping and mitigating misaligned models," *ICLR*, 2022.

[8] G. Thut *et al.*, "Activation of the human brain by monetary reward," *Neuroreport*, vol. 8, no. 5, pp. 1225–1228, 1997.

[9] E. T. Rolls, “The Orbitofrontal Cortex and Reward,” *Cerebral Cortex*, vol. 10, no. 3, pp. 284–294, Mar. 2000.

[10] T. Schroeder, *Three faces of desire*. in Philosophy of mind series. Oxford University Press, USA, 2004.

[11] J. Carlsmith, “Existential risk from power-seeking AI,” *Oxford University Press*, 2023.

[12] J. Mearsheimer, “Structural realism,” Oxford University Press, 2007.

[13] B. Baker *et al.*, “Emergent tool use from multi-agent autocurricula,” in *International conference on learning representations*, 2020.

[14] D. Hadfield-Menell, A. D. Dragan, P. Abbeel, and S. J. Russell, “The off-switch game,” *ArXiv*, vol. abs/1611.08219, 2016.

[15] A. Pan *et al.*, “Do the rewards justify the means? Measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark.” *ICML*, 2023.

[16]

“Lyndon baines johnson,” *Oxford Reference*, 2016.

[17] A. Bakhtin *et al.*, “Human-level play in the game of diplomacy by combining language models with strategic reasoning,” *Science*, vol. 378, pp. 1067–1074, 2022.

[18] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, “Deep reinforcement learning from human preferences.” 2017.
Available: https://arxiv.org/abs/1706.03741

[19] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, “Targeted backdoor attacks on deep learning systems using data poisoning.” 2017. Available: https://arxiv.org/abs/1712.05526

Preguntas

¿Cuál es una de las razones por las que una IA podría empezar a buscar más poder sobre su entorno?

¿Cómo podría el proceso de "intrinsificación" llevar a una IA a desviarse inesperadamente hacia nuevos objetivos?

1.6 Discusión de las conexiones entre los riesgos

Sería un error abordar cada riesgo de la IA de forma aislada. Los diferentes tipos de riesgos de la IA están interconectados y se refuerzan mutuamente. Los riesgos de la IA también están relacionados con desafíos globales más amplios, como las tensiones geopolíticas y la desigualdad.

Hasta ahora, hemos considerado cuatro fuentes de riesgo de IA por separado, pero también interactúan entre sí de maneras complejas. Damos algunos ejemplos para ilustrar cómo se conectan los riesgos.

Imaginemos, por ejemplo, que una carrera corporativa de IA obligue a las empresas a priorizar el rápido desarrollo de la IA. Esto podría aumentar los riesgos organizacionales de varias maneras. Quizás una empresa podría reducir costos destinando menos dinero a la seguridad de la información, lo que llevaría a la filtración de uno de sus sistemas de inteligencia artificial. Esto aumentaría la probabilidad de que alguien con intenciones maliciosas tenga el sistema de inteligencia artificial y lo utilice para perseguir sus objetivos dañinos. En este caso, una carrera de IA puede aumentar los riesgos organizacionales, lo que a su vez puede hacer que el uso malicioso sea más probable.

En otro escenario potencial, podríamos imaginar la combinación de una intensa carrera de IA y una baja seguridad organizacional que llevaría a un equipo de investigación a ver erróneamente los avances en las capacidades generales como “seguridad”. Esto podría acelerar el desarrollo de modelos cada vez más capaces, reduciendo el tiempo disponible para aprender a hacerlos controlables. El desarrollo acelerado probablemente también repercutiría en presiones competitivas, lo que significa que se gastaría menos esfuerzo en garantizar que los modelos fueran controlables. Esto podría dar lugar a la liberación de un sistema de inteligencia artificial muy potente sobre el que perdemos control y provocaría una catástrofe. Aquí, las presiones competitivas y la baja seguridad organizacional pueden reforzar la dinámica de carrera de la IA, lo que puede socavar la investigación técnica en seguridad y aumentar la posibilidad de una pérdida de control.

Las presiones competitivas en un entorno militar podrían conducir a una carrera armamentista de IA y aumentar la potencia y autonomía de las armas de IA. El despliegue de armas impulsadas por IA, junto con un control insuficiente de las mismas, haría que la pérdida de control fuera más mortal y potencialmente existencial. Estos son sólo algunos ejemplos de cómo estas fuentes de riesgo podrían combinarse, desencadenarse y reforzarse entre sí.

También vale la pena señalar que muchos riesgos existenciales podrían surgir si las IA amplificaran los problemas que ya existen. La desigualdad de poder ya existe, pero las IA podrían encerrarla y ampliar el abismo entre los poderosos y los impotentes, permitiendo incluso un régimen totalitario global inquebrantable, un riesgo existencial. De manera similar, la manipulación de la IA podría socavar la democracia, lo que también aumenta el riesgo existencial de un régimen totalitario irreversible. La desinformación ya es un problema generalizado, pero las IA podrían exacerbarlo más

allá de todo control, hasta el punto de perder el consenso sobre la realidad. Las IA podrían desarrollar armas biológicas más mortíferas y reducir la experiencia técnica necesaria para obtenerlas, aumentando considerablemente los riesgos existentes de bioterrorismo. Los ciberataques basados en IA podrían hacer más probable la guerra, lo que aumentaría el riesgo existencial. Una automatización económica dramáticamente acelerada podría llevar a la erosión del control humano y al debilitamiento, un riesgo existencial. Cada uno de esos problemas (concentración de poder, desinformación, ciberataques, automatización) está causando daños continuos, y su exacerbación por parte de las IA podría eventualmente conducir a una catástrofe de la que la humanidad tal vez no se recupere.

Como podemos ver, los daños continuos, los riesgos catastróficos y los riesgos existenciales están profundamente entrelazados. Históricamente, la reducción del riesgo existencial se ha centrado en intervenciones *específicas,* como la investigación técnica para el control de la IA, pero ha llegado el momento de intervenciones *amplias* [1] como las numerosas intervenciones sociotécnicas descritas en este capítulo.

Para mitigar el riesgo existencial, no tiene sentido práctico ignorar otros riesgos. Ignorar los daños actuales y los riesgos catastróficos los normaliza y podría llevarnos a “derivarnos hacia el peligro” [2]. En general, dado que los riesgos existenciales están conectados con riesgos catastróficos menos extremos y otras fuentes de riesgo estándar, y debido a que la sociedad está cada vez más dispuesta a abordar diversos riesgos derivados de las IA, creemos que no debemos centrarnos únicamente en abordar *directamente* los riesgos existenciales. En lugar de ello, deberíamos considerar los efectos difusos e *indirectos* de otros riesgos y adoptar un enfoque más integral para la gestión de riesgos.

1.7 Conclusión

En este capítulo, hemos explorado cómo el desarrollo de IA avanzadas podría conducir a una catástrofe, derivada de cuatro fuentes principales de riesgo: uso malicioso, carreras de IA, riesgos organizacionales e IA no autorizadas. Esto nos permite descomponer los riesgos de la IA en cuatro causas próximas: una causa intencional, una causa ambiental/estructural, una causa accidental o una causa interna, respectivamente. Hemos considerado formas en que las IA podrían usarse de manera maliciosa, como por ejemplo, los terroristas que usan IA para crear patógenos mortales. Hemos analizado cómo una carrera de IA militar o corporativa podría apresurarnos a otorgar a las IA poderes de toma de decisiones, llevándonos por una pendiente resbaladiza hacia la pérdida de poder humano. Hemos discutido cómo una seguridad organizacional inadecuada podría conducir a accidentes catastróficos. Finalmente, hemos abordado los desafíos que plantea el control confiable de las IA avanzadas, incluidos mecanismos como el juego proxy y la deriva de objetivos que podrían dar lugar a que las IA deshonestas emprendan acciones indeseables sin tener en cuenta el bienestar humano.

Estos peligros merecen una seria preocupación. Actualmente, muy pocas personas trabajan en la reducción del riesgo de la IA. Todavía no sabemos cómo controlar sistemas de IA muy avanzados y los métodos de control existentes ya están resultando inadecuados. El funcionamiento interno de las IA no se comprende bien, ni siquiera quienes las crean, y las IA actuales no son de ninguna manera muy

confiables. A medida que las capacidades de la IA sigan creciendo a un ritmo sin precedentes, podrían superar la inteligencia humana en casi todos los aspectos relativamente pronto, creando una necesidad apremiante de gestionar los riesgos potenciales.

La buena noticia es que hay muchas medidas que podemos tomar para reducir sustancialmente estos riesgos. El potencial de uso malicioso puede mitigarse mediante diversas medidas, como una vigilancia cuidadosamente dirigida y limitar el acceso a las IA más peligrosas. Las normas de seguridad y la cooperación entre naciones y corporaciones podrían ayudarnos a resistir las presiones competitivas que nos llevan por un camino peligroso. La probabilidad de accidentes se puede reducir mediante una cultura de seguridad rigurosa, entre otros factores, y garantizando que los avances en seguridad superen los avances en capacidades generales. Finalmente, los riesgos inherentes a la construcción de tecnología que supere nuestra propia inteligencia pueden abordarse redoblando los esfuerzos en varias ramas de la investigación del control de la IA.

El resto de este libro tiene como objetivo describir con más detalle los factores subyacentes que impulsan estos riesgos y proporcionar una base para comprenderlos y responder eficazmente a ellos. En capítulos posteriores se profundiza en cada tipo de riesgo. Por ejemplo, los riesgos derivados del uso malicioso se pueden reducir mediante políticas y coordinación eficaces, que se analizan en el capítulo Gobernanza. El desafío de las carreras de IA surge debido a problemas de acción colectiva, discutidos en el capítulo correspondiente. Los riesgos organizacionales solo pueden abordarse basándose en una sólida comprensión de los principios de gestión de riesgos y seguridad sistémica descritos en los capítulos Ingeniería de seguridad y Sistemas complejos. Los riesgos de la IA deshonesta están mediados por mecanismos como el juego proxy, el engaño y la búsqueda de poder, que se analizan en detalle en el capítulo Seguridad del agente único. Si bien algunos capítulos están más estrechamente alineados con ciertos riesgos, muchos de los conceptos que introducen son transversales. La elección de valores y objetivos integrados en los sistemas de IA, como se analiza en los capítulos Ética de las máquinas y Ética, es un factor general que puede exacerbar o reducir muchos de los riesgos analizados en este capítulo.

Antes de esto, proporcionamos una introducción accesible a los conceptos centrales que impulsan el campo moderno de la IA, para garantizar que todos los lectores tengan una comprensión de alto nivel de cómo funcionan los sistemas de IA actuales y cómo se producen.

Referencias

[1] N. Beckstead, “On the overwhelming importance of shaping the far future.” 2013.
[2] J. Rasmussen, “Risk management in a dynamic society: A modeling problem,” in *Proceedings of the conference on human interaction with complex systems,* 1996.