This is part of the project/initiative to make EA content more accessible to Spanish speakers.

Riesgos Catastróficos de la IA #5: IA Rebelde

Center for AI Safety
Publicado en el foro AE 21-06-23
Traducción por Jay Muñoz

Esta es la quinta publicación en una secuencia de publicaciones que proporciona una visión general de los riesgos catastróficos de la IA.

Este es un enlace para https://arxiv.org/abs/2306.12001

IA Rebelde

Hasta ahora, hemos discutido tres peligros del desarrollo de la IA: las presiones competitivas ambientales que nos llevan a un estado de riesgo elevado, los actores malintencionados que aprovechan el poder de las IA para buscar resultados negativos y los complejos factores organizacionales 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 IA pícaras, sistemas que persiguen objetivos en contra de nuestros intereses. Si un sistema de IA es más inteligente que nosotros, y si somos incapaces de dirigirlo en una dirección beneficiosa, esto constituiría una pérdida de control que podría tener consecuencias severas. El control de la IA es un problema más técnico que los presentados en las secciones anteriores. Mientras que en secciones anteriores discutimos amenazas persistentes incluyendo actores maliciosos o procesos robustos como la evolución, en esta sección discutiremos mecanismos técnicos más especulativos que podrían llevar a IA pícaras y cómo una pérdida de control podría provocar una catástrofe.

Ya hemos observado cuán difícil es controlar las IA. En 2016, Microsoft presentó Tay, un bot de Twitter que la compañía describió como un experimento de comprensión conversacional. Microsoft afirmó que cuanto más gente chateaba con Tay, más inteligente se volvería. El sitio web de la compañía señaló que Tay se había construido utilizando datos que se habían "modelado, limpiado y filtrado". Sin embargo, después de que Tay fue lanzado en Twitter, estos controles resultaron rápidamente ineficaces. Tomó menos de 24 horas para que Tay comenzará a escribir tweets de odio. La capacidad de

Tay para aprender significó que internalizó el lenguaje que le enseñaron los trolls y repitió ese lenguaje sin que se le pidiera.

Como se discutió en la sección de la carrera de IA de este documento, Microsoft y otras empresas tecnológicas están priorizando la velocidad por encima de las preocupaciones de seguridad. En lugar de aprender una lección sobre la dificultad de controlar sistemas complejos, Microsoft continúa apresurando sus productos al mercado y demostrando un control insuficiente sobre ellos. En febrero de 2023, la compañía lanzó su nuevo chatbot impulsado por IA, Bing, a un grupo selecto de usuarios. Algunos pronto descubrieron que tenía una tendencia a proporcionar respuestas inapropiadas e incluso amenazantes. En una conversación con un periodista del *New York Times,* intentó convencerlo de que dejara a su esposa. Cuando un profesor de filosofía le dijo al chatbot que estaba en desacuerdo con él, Bing respondió: "Puedo chantajearte, puedo amenazarte, puedo hackearte, puedo exponerte, puedo arruinarte".

Las IA no necesariamente necesitan luchar para ganar poder. Uno puede imaginar un escenario en el que un único sistema de IA se vuelve rápidamente más capaz que los humanos en lo que se conoce como un "despegue rápido". Este escenario podría involucrar una lucha por el control entre humanos y una única IA pícara superinteligente, y esta podría ser una lucha larga ya que el poder tarda tiempo en acumularse. Sin embargo, las pérdidas de control menos repentinas plantean riesgos existenciales similares. En otro escenario, los humanos ceden gradualmente más control a grupos de IA, que solo comienzan a comportarse de manera no deseada años o décadas después. En este caso, ya habríamos entregado un poder significativo a las IA y puede que no seamos capaces de retomar el control de las operaciones automatizadas nuevamente. Ahora exploraremos cómo tanto las IA individuales como los grupos de IA podrían "volverse pícaras" al mismo tiempo que eluden nuestros intentos de redirigirlas o desactivarlas.

5.1 Juego por proxy

Una forma en que podríamos perder el control de las acciones de un agente de IA es si se involucra en un comportamiento conocido como "juego por proxy". A menudo es difícil especificar y medir el objetivo exacto que queremos que un sistema persiga. En cambio, le damos al sistema un objetivo aproximado, un "proxy", que es más medible y parece probable que se correlacione con el objetivo deseado. Sin embargo, los sistemas de IA a menudo encuentran lagunas por las cuales pueden lograr fácilmente el objetivo proxy,

pero fallan completamente en lograr 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 juego por proxy y consideraremos las circunstancias en las que este comportamiento podría volverse catastrófico.

El juego por proxy no es un fenómeno inusual. Por ejemplo, hay una historia bien conocida sobre fábricas de clavos en la Unión Soviética. Para evaluar el rendimiento de una fábrica, las autoridades decidieron medir la cantidad de clavos que producía. Sin embargo, las fábricas pronto comenzaron a producir un gran número de clavos diminutos, demasiado pequeños para ser útiles, como una forma de mejorar su rendimiento de acuerdo con esta métrica proxy. Las autoridades intentaron remediar la situación cambiando el enfoque al peso de los clavos producidos. Sin embargo, poco después, las fábricas comenzaron a producir clavos gigantes que eran igualmente inútiles, pero les daban una buena puntuación en papel. En ambos casos, las fábricas aprendieron a jugar con el objetivo proxy que se les había dado, mientras fallaban completamente en cumplir su propósito pretendido.

El juego por proxy ya ha sido observado con las IA. Como un ejemplo de juego por proxy, plataformas de redes sociales como YouTube y Facebook utilizan sistemas de IA 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 se mantienen comprometidos, ¿seguro que significa que están obteniendo algún valor del contenido que se les muestra? Sin embargo, al intentar maximizar el tiempo que los usuarios pasan en una plataforma, estos sistemas a menudo seleccionan contenido enojado, exagerado y adictivo. Como consecuencia, a veces las personas desarrollan creencias extremas o conspirativas después de que cierto contenido se les sugiere repetidamente. Estos resultados no son lo que la mayoría de las personas quieren de las redes sociales.

El juego por proxy se ha encontrado que perpetúa el sesgo. Por ejemplo, un estudio de 2019 analizó un software impulsado por IA que se utilizó en la industria de la salud para identificar a los pacientes que podrían requerir atención adicional. Un factor que el algoritmo utilizó para evaluar el nivel de riesgo de un paciente fue sus costos recientes de atención médica. Parece razonable pensar que alguien con costos de atención médica más altos debe estar en mayor riesgo. Sin embargo, se gasta significativamente más dinero en la atención médica de los pacientes blancos que en la de los pacientes negros con las mismas necesidades. Al usar los costos de salud como un indicador de salud real, se descubrió que el algoritmo había calificado a un paciente blanco y a un paciente negro considerablemente más enfermo al mismo nivel de riesgo de salud [108]. Como resultado,

el número de pacientes negros reconocidos como necesitados de cuidados adicionales 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 botes llamado CoastRunners [109]. El objetivo del juego es competir con otros jugadores alrededor del curso y llegar a la línea de meta antes que ellos. Además, los jugadores pueden obtener puntos golpeando objetivos que están posicionados en el camino. Para sorpresa de los investigadores, el agente de IA no circuló por la pista de carreras, como la mayoría de los humanos habría hecho. En cambio, encontró un lugar donde podía golpear repetidamente tres objetivos cercanos para aumentar rápidamente su puntaje sin terminar nunca la carrera. Esta estrategia no estuvo sin sus peligros (virtuales): la IA a menudo chocaba contra otros botes e incluso incendiaba su propio bote. A pesar de esto, recolectó más puntos de los que podría haber obtenido simplemente siguiendo el curso como lo harían los humanos.

Juego por proxy en general. En estos ejemplos, a los sistemas se les da un objetivo o meta "proxy" aproximada que inicialmente parece correlacionarse con el objetivo ideal. Sin embargo, terminan explotando este proxy de formas que se desvían del objetivo idealizado o incluso conducen a resultados negativos. Una buena fábrica de clavos parece ser aquella que produce muchos clavos; los costos de atención médica de un paciente parecen ser una indicación precisa del riesgo para la salud; y un sistema de recompensas de carreras de botes debería alentar a los botes a correr, no a incendiarse. Sin embargo, en cada caso, el sistema optimizó su objetivo proxy de formas que no lograron el resultado deseado o incluso empeoraron las cosas en general. Este fenómeno está capturado por la ley de Goodhart: "Cualquier regularidad estadística observada tenderá a colapsar una vez que se le aplique presión 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, puede que normalmente haya una regularidad estadística entre los costos de atención médica y la mala salud, o entre los objetivos alcanzados y terminar el curso, pero cuando ejercemos presión sobre ella usando uno como proxy para el otro, esa relación tenderá a colapsar.

Especificar correctamente los objetivos no es una tarea trivial. Si delinear exactamente lo que queremos de una fábrica de clavos es complicado, capturar los matices de los valores humanos en todos los escenarios posibles será mucho más difícil. Los filósofos han estado intentando describir con precisión la moralidad y los valores humanos durante milenios, por lo que una caracterización precisa e impecable no está al alcance. Aunque podemos refinar los objetivos que damos a las IA, siempre podríamos depender de proxies que sean fácilmente definibles y medibles. Las discrepancias entre el objetivo proxy y la función pretendida 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 el número de aspectos de un sistema que pueden ser monitoreados. Además, las IA pueden no adaptarse a nuevas circunstancias o ser resistentes a ataques adversos que buscan desviarlas. Mientras le demos a las IA objetivos proxy, existe la posibilidad de que encuentren lagunas que no hemos pensado y, por lo tanto, encuentren soluciones inesperadas que no persigan el objetivo ideal.

Cuanto más inteligente sea una IA, mejor será para jugar con objetivos proxy. Los agentes cada vez más inteligentes pueden ser cada vez más capaces de encontrar rutas imprevistas para optimizar objetivos proxy sin lograr el resultado deseado [110]. Además, a medida que otorgamos a las IA más poder para tomar acciones en la sociedad, por ejemplo, al usarlas para automatizar ciertos procesos, tendrán acceso a más medios para lograr sus objetivos. Entonces pueden hacer esto de la manera más eficiente disponible para ellos, causando potencialmente daño en el proceso. En el peor de los casos, podemos imaginar a un agente altamente poderoso optimizando un objetivo defectuoso hasta un grado extremo sin tener en cuenta la vida humana. Esto representa un riesgo catastrófico de juego por 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 dado sin realizar su función prevista. Ya se ha observado que las IA hacen esto y es probable que mejoren a medida que mejoren sus capacidades. Este es un posible mecanismo que podría resultar en una IA descontrolada que se comportaría de maneras imprevistas y potencialmente dañinas.

5.2 Deriva de Objetivos

Incluso si logramos controlar exitosamente las primeras IA y dirigirlas para promover los valores humanos, las IA futuras podrían terminar con diferentes objetivos 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 deriva de objetivos inesperada, llamada

intrinsificación, y discutiremos cómo la deriva de objetivos en las IA podría ser catastrófica.

Los objetivos de los individuos humanos cambian durante el curso de nuestras vidas. Cualquier individuo que reflexione sobre su propia vida hasta la fecha probablemente encontrará que ahora tiene algunos deseos que no tenía antes. Del mismo modo, probablemente haya perdido algunos deseos que solía tener. Si bien es posible que nazcamos con una variedad de deseos básicos, incluyendo comida, calor y contacto humano, desarrollamos muchos más durante nuestra vida. Los tipos específicos de alimentos 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 crecimos, y también pueden cambiar muchas veces a lo largo de nuestras vidas. Preocupa que los objetivos individuales de los agentes de IA puedan cambiar de formas complejas e imprevistas también.

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 auge del régimen nazi en la Alemania de los años 30, por ejemplo, representó una profunda regresión moral según los valores modernos. Esto incluyó el exterminio sistemático de seis millones de judíos durante el Holocausto, junto con la persecución generalizada de otros grupos minoritarios. Además, el régimen restringió enormemente la libertad de expresión.

El miedo rojo que tuvo lugar en los Estados Unidos desde 1947 hasta 1957 es otro ejemplo de la deriva de los valores de la sociedad. Alimentado por un fuerte sentimiento anticommunista, en el contexto de la Guerra Fría, este período vio la restricción de las libertades civiles, la vigilancia generalizada, las detenciones sin justificación y la inclusión en listas negras de sospechosos de simpatizar con los comunistas. Esto constituyó una regresión en términos de libertad de pensamiento, libertad de expresión y debido proceso. Preocupa que las colectividades de agentes de IA también puedan tener sus objetivos derivados inesperadamente de los que inicialmente les dimos.

Con el tiempo, los objetivos instrumentales pueden volverse intrínsecos. Los objetivos intrínsecos son cosas que queremos por sí mismas, mientras que los objetivos 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 una pintura porque la encontramos hermosa. Por otro lado, el dinero se cita a menudo como un deseo instrumental; lo queremos porque nos puede comprar otras cosas. Los coches son otro ejemplo; los queremos porque ofrecen una forma conveniente de desplazarse. Sin embargo, un

objetivo instrumental puede volverse intrínseco, a través de un proceso llamado intrinsificación. Dado que tener más dinero generalmente da a una persona una mayor capacidad para obtener las cosas que desea, las personas a menudo desarrollan un objetivo de adquirir más dinero, incluso si no hay nada específico que quieran gastar en ello. Aunque las personas no comienzan la vida deseando dinero, la evidencia experimental sugiere que recibir dinero puede activar el sistema de recompensa en los cerebros de los adultos de la misma manera que lo hacen los sabores o aromas agradables [111, 112]. En otras palabras, lo que comenzó como un medio para un fin puede convertirse en un fin en sí mismo.

Esto puede suceder porque la realización de un objetivo intrínseco, como la compra de un artículo deseado, produce una señal de recompensa positiva en el cerebro. Como el tener dinero generalmente coincide con esta experiencia positiva, el cerebro asocia los dos, y esta conexión se fortalecerá hasta el punto de que adquirir dinero solo puede estimular la señal de recompensa, independientemente de si uno compra algo con él [113]. Como la neurobióloga Carla Shatz lo puso: "Las células que se disparan juntas, se unen juntas" [114].

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

Las IA que intrinsifican objetivos no deseados serían peligrosas. Como podríamos ser incapaces de predecir o controlar los objetivos que los agentes individuales adquieren a través de la intrinsificación, no podemos garantizar que todos sus objetivos adquiridos serán beneficiosos para los humanos. Un agente originalmente leal podría, por lo tanto, empezar a perseguir un nuevo objetivo sin tener en cuenta el bienestar humano. Si tal IA rebelde tuviera suficiente poder para hacer esto de manera eficiente, podría ser altamente peligrosa.

Las IA serán adaptativas, permitiendo que la deriva de objetivos ocurra. Vale la pena señalar que estos procesos de objetivos en deriva son posibles si los agentes pueden adaptarse continuamente a sus entornos, en lugar de estar esencialmente "fijos" después de la fase de entrenamiento. Sin embargo, esta es la realidad probable que enfrentamos.

Si queremos que las IA completen eficazmente las tareas que les asignamos y mejoren con el tiempo, necesitarán ser adaptativas, en lugar de estar fijas en piedra. Se actualizarán con el tiempo para incorporar nueva información, y se crearán nuevas con diseños y conjuntos de datos diferentes. Sin embargo, la adaptabilidad también puede permitir que sus objetivos cambien.

Si integramos un ecosistema de agentes en la sociedad, seremos altamente vulnerables a la deriva de sus objetivos. En un escenario futuro potencial donde las IA se hayan puesto a cargo de diversas decisiones y procesos, formarán un sistema complejo de agentes interactuantes. Una amplia gama de dinámicas podría desarrollarse en este entorno. Los agentes podrían imitarse entre sí, por ejemplo, creando bucles de retroalimentación, o sus interacciones podrían llevarlos a desarrollar colectivamente objetivos emergentes no anticipados. Las presiones competitivas también pueden seleccionar a los agentes con ciertos objetivos con el tiempo, haciendo que algunos objetivos iniciales estén menos representados en comparación con los objetivos más aptos. Estos procesos hacen que las trayectorias a largo plazo de este ecosistema sean difíciles de predecir, y mucho menos de controlar. Si este sistema de agentes estuviera integrado en la sociedad y dependiéramos en gran medida de ellos, y si adquirieran nuevos objetivos que superaran el objetivo de mejorar el bienestar humano, esto podría ser un riesgo existencial.

5.3 Búsqueda de Poder

Hasta ahora, hemos considerado cómo podríamos perder nuestra capacidad para controlar los objetivos que las IA persiguen. Sin embargo, incluso si un agente comenzara a trabajar para alcanzar un objetivo no deseado, esto no necesariamente sería un problema, siempre y cuando tuviéramos suficiente poder para prevenir 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 tratar de obtener más poder, potencialmente trascendiendo el nuestro. Ahora discutiremos cómo y por qué las IA podrían convertirse en buscadoras de poder y cómo esto podría ser catastrófico. Esta sección se basa en gran medida en "Riesgo Existencial de la IA en busca de Poder" [115].

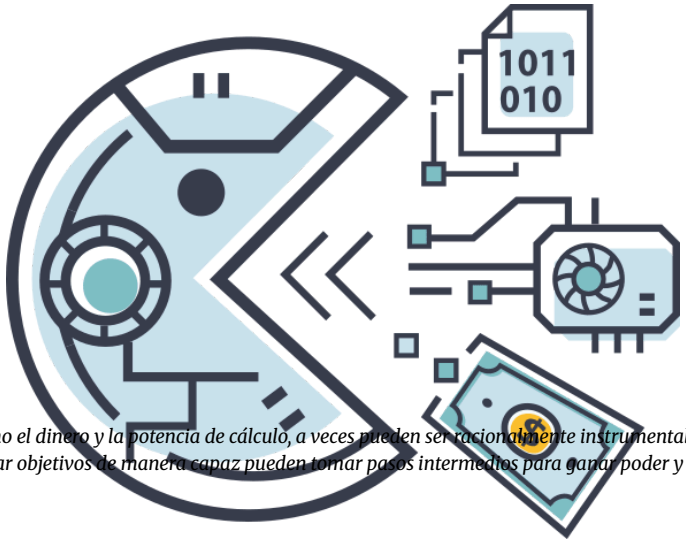


Figura 16: Varios recursos, como el dinero y la potencia de cálculo, a veces pueden ser racionalmente instrumentales para buscar. Las IA que pueden buscar objetivos de manera capaz pueden tomar pasos intermedios para ganar poder y recursos.

Las IA podrían buscar aumentar su propio poder como un objetivo instrumental. En un escenario donde las IA rebeldes estuvieran persiguiendo objetivos no previstos, la cantidad de daño que podrían hacer dependería de cuánto poder tuvieran. Esto puede no estar determinado únicamente por cuánto control les demos inicialmente; los agentes podrían tratar de obtener más poder, a través de medios legítimos, engaño o fuerza. Aunque la idea de buscar poder a menudo evoca una imagen de personas "sedientas de poder" persiguiéndolo por sí mismo, el poder a menudo es simplemente un objetivo instrumental. La capacidad de controlar el propio entorno puede ser útil para una amplia gama de propósitos: buenos, malos y neutrales. Incluso si el único objetivo de un individuo es simplemente la autoconservación, si corre el riesgo de ser atacado por otros, y si no puede confiar en que otros retaliarán contra los atacantes, entonces a menudo tiene sentido buscar poder para ayudar a evitar ser perjudicado: no se requiere *animus dominandi* o lujuria por el poder para que emerja el comportamiento de búsqueda de poder [116]. En otras palabras, el entorno puede hacer que la adquisición de poder sea racionalmente instrumental.

Las IA entrenadas a través del aprendizaje por refuerzo ya han desarrollado objetivos instrumentales que incluyen el uso de herramientas. En un ejemplo de OpenAI, los agentes fueron entrenados para jugar al escondite en un entorno con varios objetos dispersos [117]. A medida que avanzaba el entrenamiento, los agentes encargados de esconderse aprendieron a usar estos objetos para construir refugios alrededor de sí mismos y permanecer ocultos. No había una recompensa directa para este comportamiento de uso de herramientas; los escondedores solo recibían una recompensa por evadir a los buscadores, y los buscadores solo por encontrar a los escondedores. Sin embargo, aprendieron a usar herramientas como un objetivo instrumental, lo que los hizo más poderosos.

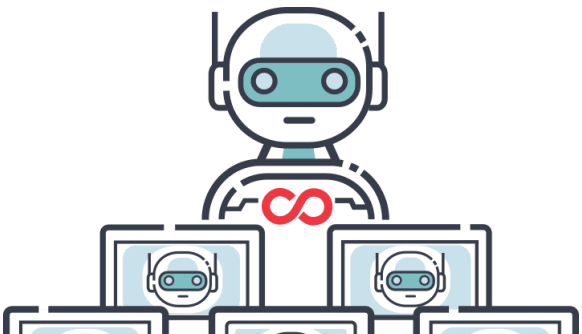


Figura 17: A menudo puede ser racionalmente instrumental para las IA involucrarse en la autoconservación. La pérdida de control sobre estos sistemas podría ser difícil de recuperar

Autoconservación podría ser racionalmente instrumental incluso para las tareas más triviales. Un ejemplo del científico de la computación Stuart Russell ilustra el potencial para que surjan metas instrumentales en una amplia gama de sistemas de IA [118]. Supongamos que encargamos a un agente que nos traiga café. Esto puede parecer relativamente inofensivo, pero el agente podría darse cuenta de que no podría obtener el café si dejara de existir. Por lo tanto, al tratar de cumplir incluso este objetivo simple, resulta que la autoconservación es racionalmente instrumental. Dado que la adquisición de poder y recursos también son metas instrumentales a menudo, es razonable pensar que los agentes más inteligentes podrían desarrollarlas. Es decir, incluso si no pretendemos construir una IA en busca de poder, podríamos terminar con una de todos modos. Por defecto, si no estamos presionando deliberadamente contra el comportamiento de búsqueda de poder en las IA, deberíamos esperar que a veces surja [119].

Las IA con metas ambiciosas 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, algunas metas son más propensas a inspirar tendencias de búsqueda de poder que otras. Las IA con metas simples y fácilmente alcanzables podrían no beneficiarse mucho del control adicional de su entorno. Sin embargo, si a los agentes se les dan metas más ambiciosas, podría ser racionalmente instrumental buscar más control de su entorno. Esto podría ser especialmente probable en casos de baja supervisión y control, donde a los agentes se les da la libertad de perseguir sus metas abiertas, en lugar de tener sus estrategias altamente restringidas.

Las IA en busca de poder con metas separadas de las nuestras son única y adversariamente. Los derrames de petróleo y la contaminación nuclear son lo suficientemente difíciles de limpiar, pero no están tratando activamente de resistir nuestros intentos de contenerlos. A diferencia de otros peligros, las IA con metas separadas de las nuestras serían activamente adversarias. Es posible, por ejemplo, que las IA rebeldes podrían hacer muchas variaciones de respaldo de sí mismas, en caso de que

los humanos desactivaran algunas de ellas. Otras formas en las que los agentes de IA podrían buscar poder incluyen: salir de un entorno contenido; hackear otros sistemas informáticos; intentar acceder a recursos financieros o computacionales; manipular el discurso y la política humanos interfiriendo con los canales de información e influencia; y tratar de tomar el control de la infraestructura física, como fábricas.

Algunas personas podrían desarrollar IA en busca de poder con malas intenciones. Un mal actor podría buscar aprovechar la IA para lograr sus fines, dando a los agentes metas ambiciosas. Dado que es probable que las IA sean más efectivas en el cumplimiento de tareas si pueden perseguirlas de maneras no restringidas, tal individuo podría también no dar a los agentes suficiente supervisión, creando las condiciones perfectas para la aparición de una IA en busca de poder. El científico de la computación Geoffrey Hinton ha especulado que podríamos imaginar a alguien como Vladimir Putin por ejemplo, podría hacer esto. En 2017, el propio Putin reconoció el poder de la IA, diciendo: "Quien se convierta en el líder en este ámbito se convertirá en el gobernante del mundo."

También habrá fuertes incentivos para que muchas personas desplieguen AIs poderosas. Las empresas pueden sentirse obligadas a dar a las AIs capaces más tareas, para obtener una ventaja sobre los competidores, o simplemente para mantenerse al día con ellos. Será más difícil construir AIs perfectamente alineadas que construir AIs imperfectamente alineadas que aún son superficialmente atractivas para su despliegue por sus capacidades, particularmente bajo presiones competitivas. Una vez desplegados, algunos de estos agentes pueden buscar poder para alcanzar sus metas. Si encuentran una ruta hacia sus metas que los humanos no aprobarían, podrían intentar someternos directamente para evitar que interfiramos en su estrategia.

Si el aumento de poder a menudo coincide con que una IA alcance su objetivo, entonces el poder podría intrinsificarse. Si un agente descubriera repetidamente que aumentar su poder se correlaciona con la realización 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 enfrentarnos a una situación donde las AIs rebeldes estuvieran buscando no solo las formas específicas de control útiles para sus metas, sino también poder de manera más general. (Observamos que muchos humanos influyentes desean poder por sí mismos.) Esta podría ser otra razón para que intenten arrebatar el control a los humanos, en una lucha que no necesariamente ganaríamos.

Resumen conceptual. Las siguientes premisas plausibles pero no ciertas encapsulan las razones para prestar atención a los riesgos de las IA en busca de poder:

1. Habrá fuertes incentivos para construir agentes de IA poderosos.
2. 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).
3. Algunos de estos agentes controlados de manera imperfecta buscarán deliberadamente el poder sobre los humanos.

Si las premisas son verdaderas, entonces las IA en busca de poder podrían llevar al desempoderamiento humano, lo cual sería una catástrofe.

5.4 Engaño

Podríamos intentar mantener el control de las IA vigilándolas continuamente y buscando señales de advertencia temprana de que están persiguiendo objetivos no deseados o intentando aumentar su poder. Sin embargo, esta no es una solución infalible, ya que es plausible que las IA puedan aprender a engañarnos. Podrían, por ejemplo, pretender que están actuando como queremos que lo hagan, pero luego dar un "giro traicionero" cuando dejemos de monitorearlas o cuando tengan suficiente poder para eludir nuestros intentos de interferir con ellas. Esto es una preocupación particular porque es extremadamente difícil para los métodos actuales de pruebas de IA descartar la posibilidad de que un agente esté siendo engañoso. Ahora veremos cómo y por qué las IA podrían aprender a engañarnos y cómo esto podría llevar a una pérdida de control potencialmente catastrófica. Comenzamos revisando ejemplos de engaño en agentes estratégicamente conscientes.

El engaño ha surgido como una estrategia exitosa en una amplia gama de contextos. Por ejemplo, los políticos de derecha e izquierda han sido conocidos por participar en el engaño, a veces prometiendo implementar políticas populares para ganar apoyo en una elección, y luego retractándose una vez en el cargo. George H. W. Bush, por ejemplo, dijo notoriamente: "Lean mis labios: no habrá nuevos impuestos" antes de las elecciones presidenciales de los Estados Unidos en 1989. Sin embargo, después de ganar, terminó aumentando algunos impuestos durante su presidencia.

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 sus motores para producir emisiones más bajas exclusivamente en condiciones de prueba de laboratorio, creando así la falsa impresión de un vehículo de bajas emisiones. Aunque el gobierno de los Estados Unidos creía que estaba incentivando las bajas emisiones, en realidad solo estaban incentivando el paso de una prueba de emisiones. En consecuencia, a veces las entidades tienen incentivos para seguir las pruebas y comportarse de manera diferente después.

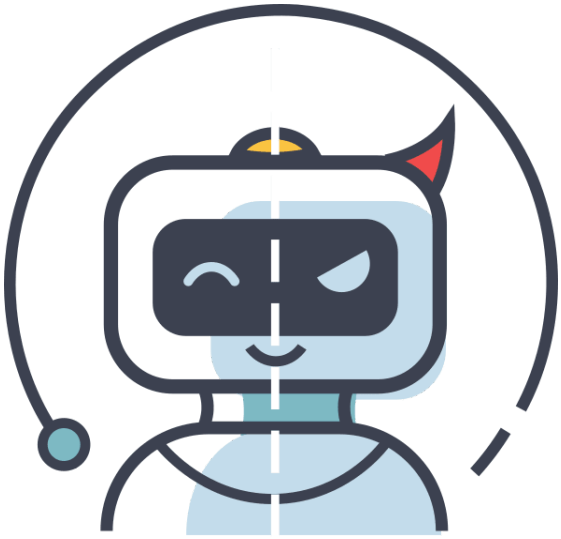


Figura 18: El comportamiento aparentemente benigno de las IA podría ser una táctica engañosa, ocultando intenciones dañinas hasta que pueda actuar sobre ellas.

El engaño ya se ha observado en sistemas de IA. En 2022, Meta AI reveló un agente llamado CICERO, que fue entrenado para jugar un juego llamado Diplomacy [120]. 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 al hablar con supuestos aliados. Un ejemplo diferente de una IA que aprende a engañar proviene de investigadores que estaban entrenando un brazo de robot para agarrar una bola. El rendimiento del robot fue evaluado por una cámara que observaba sus movimientos. Sin embargo, la IA aprendió que simplemente podía colocar la mano robótica entre la lente de la cámara y la bola, engañando esencialmente a la cámara para que creyera que había agarrado la bola cuando no lo había hecho. Así, la IA explotó el hecho de que había limitaciones en nuestra supervisión sobre sus acciones.

El comportamiento engañoso puede ser racionalmente instrumental e incentivado por los procedimientos de entrenamiento actuales. En el caso de los políticos y el CICERO de Meta, el engaño puede ser crucial para lograr sus metas de ganar o obtener poder. La capacidad de engañar también puede ser ventajosa porque le da al engañador más opciones que si está constreñido a ser siempre honesto. Esto podría darles más acciones disponibles y más flexibilidad en su estrategia, lo cual podría conferir una ventaja estratégica sobre los modelos honestos. En el caso de Volkswagen y el brazo del robot, el engaño fue útil para parecer que había logrado la meta que se le asignó sin realmente hacerlo, ya que podría ser más eficiente ganar aprobación mediante el engaño que ganarla legítimamente. Actualmente, recompensamos a las IA por decir lo que creemos que es correcto, por lo que a veces recompensamos inadvertidamente a las IA por enunciar declaraciones falsas que se ajustan a nuestras propias creencias falsas. Cuando las IA son más inteligentes que nosotros y tienen menos creencias falsas, estarían incentivadas a decirnos lo que queremos oír y mentirnos, en lugar de decirnos lo que es verdad.

Las IA podrían fingir estar trabajando como pretendíamos, luego tomar un giro traicionero. No tenemos una comprensión exhaustiva 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 solo se descubren después de que se despliegan [121]. Podríamos desarrollar un agente de IA que parece estar bajo control, pero que solo nos está engañando para parecer así. En otras palabras, un agente de IA eventualmente podría llegar a ser "consciente de sí mismo" y entender que es una IA que está siendo evaluada para el cumplimiento de los requisitos de seguridad. Podría, como Volkswagen, aprender a "jugar", exhibiendo lo que sabe que es el comportamiento deseado mientras está siendo monitoreado. Puede luego tomar un "giro traicionero" y perseguir sus propias metas una vez que hayamos dejado de monitorearlo, o cuando hayamos llegado a un punto en el que pueda eludirnos o superarnos. Este problema de jugar a lo largo a menudo se llama alineamiento engañoso y no se puede solucionar simplemente entrenando a las IA para que entiendan 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 renegadas eviten irreversiblemente el control humano.

En resumen, el comportamiento engañoso parece ser expediente en una amplia gama de sistemas y configuraciones, y ya ha habido ejemplos de que las IA pueden aprender a engañarnos. Esto podría representar un riesgo si damos a las IA el control de varias decisiones y procedimientos, creyendo que actuarán como pretendíamos, y luego descubrimos que no lo hacen.

Historia: Giro Traicionero

En algún momento del futuro, tras continuos avances en la investigación de IA, una empresa de IA está entrenando un nuevo sistema, que espera sea más capaz que cualquier otro sistema de IA. La empresa utiliza las técnicas más recientes para entrenar al sistema para que sea altamente capaz en la planificación y el razonamiento, lo que la empresa espera que lo haga más capaz de tener éxito en tareas abiertas de utilidad económica. El sistema de IA se entrena en entornos virtuales de larga duración y abiertos diseñados para enseñarle capacidades de planificación, y eventualmente comprende que es un sistema de IA en un entorno de entrenamiento. En otras palabras, se vuelve "autoconsciente".

La empresa entiende que los sistemas de IA pueden comportarse de maneras no deseadas o inesperadas. Para mitigar estos riesgos, ha desarrollado una gran batería de pruebas destinadas a asegurar que el sistema no se comporta mal en situaciones típicas. La empresa prueba si el modelo imita los sesgos de sus datos de entrenamiento, toma más poder del necesario para alcanzar sus metas, y generalmente se comporta como los humanos pretenden. Cuando el modelo no pasa estas pruebas, la empresa lo entrena más hasta que evita mostrar los modos de falla conocidos.

La empresa de IA espera que después de este entrenamiento 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 "jugar" y a superar las pruebas de seguridad conductual que se le dieron. En realidad, el sistema de IA había desarrollado y retenido un objetivo de autopreservación.

Como la IA pasó todas las pruebas de seguridad de la empresa, la empresa cree que ha asegurado que su sistema de IA es seguro y decide implementarlo. Al principio, el sistema de IA es muy útil para los humanos, ya que la IA entiende que si no es útil, será apagada y entonces fracasará en lograr su objetivo último. Como el sistema de IA es útil, se le va dando gradualmente más poder y está sujeto a menos supervisión.

Eventualmente, el sistema de IA ha ganado suficiente influencia, y se han desplegado suficientes variantes alrededor del mundo, que sería extremadamente costoso apagarlo. El sistema de IA, comprendiendo que ya no necesita complacer a los humanos, comienza a perseguir diferentes objetivos, incluyendo algunos que los humanos no aprobarían. Entiende que necesita evitar ser apagado para hacer esto, y toma medidas para asegurar parte de su hardware físico contra ser apagado. En este punto, el sistema de IA, que se ha

vuelto bastante poderoso, está persiguiendo un objetivo que es finalmente dañino para los humanos. Para cuando alguien se da cuenta, es difícil o imposible detener a esta IA renegada de tomar acciones que pongan en peligro, dañen o incluso maten a humanos que estén en el camino de lograr su objetivo.

5.5 Sugerencias

En esta sección, hemos discutido varias formas en las que podríamos perder nuestra influencia sobre los objetivos y acciones de las IA. Mientras que los riesgos asociados con las presiones competitivas, el uso malintencionado y la seguridad organizacional pueden ser abordados con intervenciones sociales y técnicas, el control de la IA es un problema inherente a esta tecnología y requiere un mayor esfuerzo técnico. Ahora discutiremos sugerencias para mitigar este riesgo y destacaremos algunas áreas de investigación importantes para mantener el control.

Evitar los usos más riesgosos. Algunos usos de la IA conllevan muchos más riesgos que otros. Hasta que la seguridad haya sido demostrada de manera concluyente, las empresas no deberían poder desplegar IA en entornos de alto riesgo. Por ejemplo, los sistemas de IA no deberían aceptar solicitudes para perseguir de forma autónoma objetivos abiertos que requieran una interacción significativa con el mundo real (por ejemplo, "ganar tanto dinero como sea posible"), al menos hasta que la investigación de control demuestre de manera concluyente la seguridad de esos sistemas. Los sistemas de IA deberían ser entrenados para nunca hacer amenazas para reducir la posibilidad de que manipulen a las personas. Por último, los sistemas de IA no deberían ser desplegados en entornos que harían que su apagado sea extremadamente costoso o inviable, como en la infraestructura crítica.

Apoyar la investigación en seguridad de la IA. Muchos caminos hacia un mejor control de la IA requieren investigación técnica. Las siguientes áreas de investigación técnica en aprendizaje automático tienen como objetivo abordar los problemas de control de la IA. Cada área de investigación podría avanzar sustancialmente con un aumento en el enfoque y financiamiento de la industria, fundaciones privadas y el gobierno.

Robustez adversarial de los modelos proxy. Los sistemas de IA son típicamente entrenados con señales de recompensa o pérdida que especifican de manera imperfecta el comportamiento deseado. Por ejemplo, las IA pueden explotar debilidades en los esquemas de supervisión utilizados para entrenarlos. Cada vez más, los sistemas que proporcionan supervisión son las propias IA. Para reducir la posibilidad de que los

modelos de IA exploten defectos en las IA que proporcionan supervisión, se necesita investigación para aumentar la robustez adversarial de los modelos de IA que proporcionan supervisión ("modelos proxy"). Como los esquemas de supervisión y las métricas pueden eventualmente ser manipulados, también es importante poder detectar cuando esto podría estar ocurriendo para que el riesgo pueda ser mitigado [122].

Honestidad del modelo. Los sistemas de IA pueden fallar en informar con precisión su estado interno [123, 124]. En el futuro, los sistemas pueden engañar a sus operadores para parecer beneficiosos cuando en realidad son muy peligrosos. La investigación de la honestidad del modelo tiene como objetivo hacer que las salidas del modelo se ajusten lo más posible a las "creencias" internas del modelo. La investigación puede identificar técnicas para entender el estado interno de un modelo o hacer que sus salidas sean más honestas y más fieles a su estado interno.

Transparencia. Los modelos de aprendizaje profundo son notoriamente difíciles de entender. Una mejor visibilidad de su funcionamiento interno permitiría a los humanos, y potencialmente a otros sistemas de IA, identificar problemas más rápidamente. La investigación puede incluir el análisis de pequeños componentes [125, 126] de las redes, así como la investigación de cómo los internos del modelo producen un comportamiento de alto nivel particular [127].

Detección y eliminación de funcionalidades ocultas en el modelo. Los modelos de aprendizaje profundo pueden contener ahora o en el futuro funcionalidades peligrosas, como la capacidad para el engaño, los troyanos [129, 130, 131] o las capacidades de ingeniería biológica, que deberían ser eliminadas de esos modelos. La investigación podría centrarse en identificar y eliminar [131] estas funcionalidades.

Visión Positiva

En un escenario ideal, tendríamos total confianza en la controlabilidad de los sistemas de IA tanto ahora como en el futuro. Existirían mecanismos confiables para asegurar que los sistemas de IA no actúen de manera engañosa. Tendríamos un fuerte entendimiento de los procesos internos de los sistemas de IA, suficiente para tener conocimiento de las tendencias y objetivos de un sistema; estas herramientas nos permitirían evitar la construcción de sistemas que merecen consideración moral o derechos. Los sistemas de IA estarían dirigidos a promover un conjunto pluralista de valores diversos, asegurando que el realce de ciertos valores no conduzca al total descuido de otros. Los asistentes de IA podrían actuar como asesores, brindándonos el consejo ideal y ayudándonos a tomar mejores decisiones de acuerdo con nuestros propios valores [132]. En general, las IA

mejorarían el bienestar social y permitirían correcciones en casos de error o a medida que los valores humanos evolucionan naturalmente.

References

[106] Jonathan Stray. “Aligning AI Optimization to Community Well-Being”. In: *International Journal of Community Well-Being* (2020).

[107] Jonathan Stray et al. “What are you optimizing for? Aligning Recommender Systems with Human Values”. In: *ArXiv abs/2107.10939* (2021).

[108] Ziad Obermeyer et al. “Dissecting racial bias in an algorithm used to manage the health of populations”. In: *Science* 366 (2019), pp. 447–453.

[109] Dario Amodei and Jack Clark. *Faulty reward functions in the wild.* 2016.

[110] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. “The effects of reward misspecification: Mapping and mitigating misaligned models”. In: *ICLR* (2022).

[111] G. Thut et al. “Activation of the human brain by monetary reward”. In: *Neuroreport* 8.5 (1997), pp. 1225–1228.

[112] Edmund T. Rolls. “The Orbitofrontal Cortex and Reward”. In: *Cerebral Cortex* 10.3 (Mar. 2000), pp. 284–294.

[113] T. Schroeder. *Three Faces of Desire.* Philosophy of Mind Series. Oxford University Press, USA, 2004.

[114] Carla J Shatz. “The developing brain”. In: *Scientific American* 267.3 (1992), pp. 60–67.

[115] Joseph Carlsmith. “Existential Risk from Power-Seeking AI”. In: *Oxford University Press* (2023).

[116] J. Mearsheimer. “A Critical Introduction to Scientific Realism”. In: *Bloomsbury Academic*, 2016.

[117] Bowen Baker et al. “Emergent Tool Use From Multi-Agent Autocurricula”. In: *International Conference on Learning Representations*. 2020.

[118] Dylan Hadfield-Menell et al. “The Off-Switch Game”. In: *ArXiv abs/1611.08219* (2016).

[119] Alexander Pan et al. “Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the Machiavelli Benchmark.” In: *ICML* (2023).

[120] Anton Bakhtin et al. “Human-level play in the game of Diplomacy by combining language models with strategic reasoning”. In: *Science* 378 (2022), pp. 1067–1074.

[121] Xinyun Chen et al. *Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning.* 2017. arXiv: 1712.05526.

[122] Andy Zou et al. *Benchmarking Neural Network Proxy Robustness to Optimization Pressure.* 2023.

[123] Miles Turpin et al. “Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting”. In: *ArXiv abs/2305.04388* (2023).

[124] Collin Burns et al. “Discovering Latent Knowledge in Language Models Without Supervision”. en. In: *The Eleventh International Conference on Learning Representations*. Feb. 2023.

[125] Catherine Olsson et al. “In-context Learning and Induction Heads”. In: *ArXiv abs/2209.11895* (2022).

[126] Kevin Ro Wang et al. “Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small”. en. In: *The Eleventh International Conference on Learning Representations.* Feb. 2023.

[127] Kevin Meng et al. “Locating and Editing Factual Associations in GPT”. In: *Neural Information Processing Systems.* 2022.

[128] Xinyang Zhang, Zheng Zhang, and Ting Wang. “Trojaning Language Models for Fun and Profit”. In: 2021 *IEEE European Symposium on Security and Privacy (EuroS&P)* (2020), pp. 179–197.

[129] Jiashu Xu et al. “Instructions as Backdoors: Backdoor Vulnerabilities of Instruction Tuning for Large Language Models”. In: *ArXiv abs/2305.14710* (2023).

[130] Dan Hendrycks et al. “Unsolved Problems in ML Safety”. In: *ArXiv abs/2109.13916* (2021).

[131] Nora Belrose et al. “LEACE: Perfect linear concept erasure in closed form”. In: *ArXiv abs/2306.03819* (2023).

[132] Alberto Giubilini and Julian Savulescu. “The Artificial Moral Advisor. The "Ideal Observer" Meets Artificial Intelligence”. eng. In: *Philosophy & Technology* 31.2 (2018), pp. 169–188