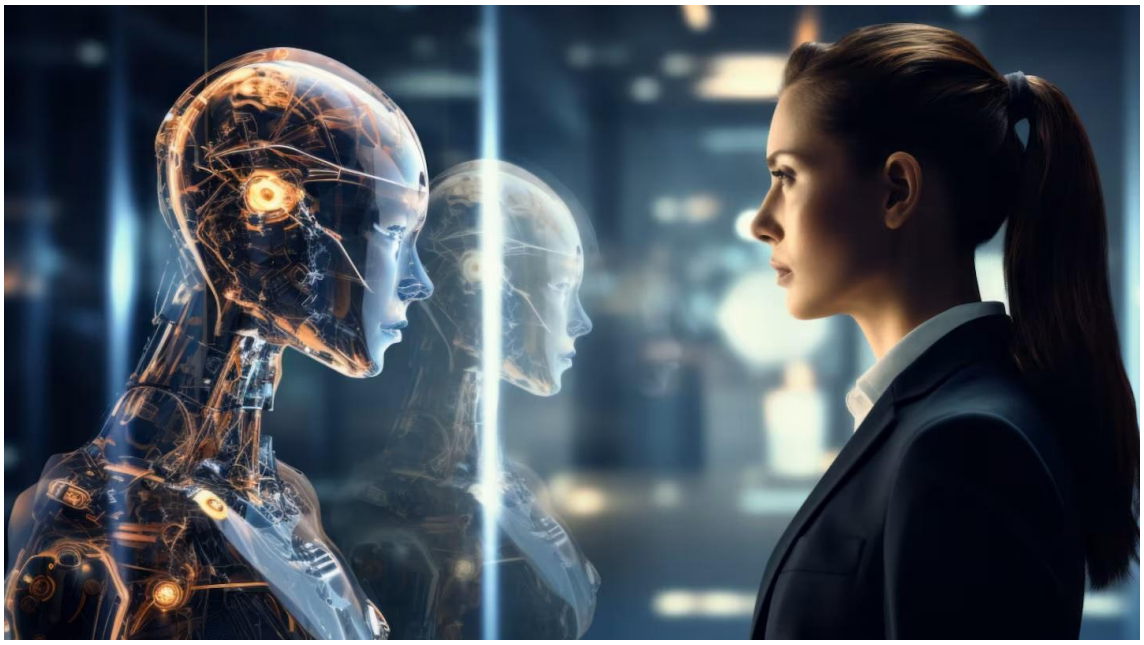

Dario Amodei

La adolescencia de la tecnología

Contenido

La adolescencia de la tecnología

The Adolescence of Technology

Confrontar y superar los riesgos de una IA poderosa
Enero 2026

Hay una escena en la versión cinematográfica del libro de Carl Sagan *Contact* donde el personaje principal, un astrónomo que ha detectado la primera señal de radio de una civilización alienígena, está siendo considerado para el papel del representante de la humanidad para conocer a los alienígenas. El panel internacional que la entrevista pregunta: "Si pudieras hacer [a los extraterrestres] solo una pregunta, ¿cuál sería?" Su respuesta es: "Yo les preguntaba: '¿Cómo lo

hiciste? ¿Cómo evolucionaste, cómo sobreviviste a esta adolescencia tecnológica sin destruirte a ti mismo?" Cuando pienso en dónde está la humanidad ahora con la IA, sobre lo que estamos en la cúspide, mi mente sigue volviendo a esa escena, porque la pregunta es muy adecuada para nuestra situación actual, y desearía tener la respuesta de los extraterrestres para guiarnos. Creo que estamos entrando en un rito de paso, tanto turbulento como inevitable, que pondrá a prueba quiénes somos como especie. La humanidad está a punto de recibir un poder casi inimaginable, y está profundamente claro si nuestros sistemas sociales, políticos y tecnológicos poseen la madurez para ejercerlo.

En mi ensayo *Máquinas de la Gracia Amorosa*, traté de exponer el sueño de una civilización que había llegado a la edad adulta, donde se habían abordado los riesgos y se aplicaba una poderosa IA con habilidad y compasión para elevar la calidad de vida de todos. Sugerí que la IA podría contribuir a enormes avances en biología, neurociencia, desarrollo económico, paz global y trabajo y significado. Sentí que era importante dar a la gente algo inspirador por la que luchar, una tarea en la que tanto los aceleracionistas de la IA como los defensores de la seguridad de la IA parecían, curiosamente, haber fracasado. Pero en este ensayo actual, quiero enfrentar el rito de la iniciación en sí: trazar los riesgos que estamos a punto de enfrentar y tratar de comenzar a hacer un plan de batalla para derrotarlos. Creo profundamente en nuestra capacidad de prevalecer, en el espíritu de la humanidad y su nobleza, pero debemos enfrentar la situación de manera directa y sin ilusiones.

Al igual que con hablar de los beneficios, creo que es importante discutir los riesgos de una manera cuidadosa y bien considerada. En particular, creo que es fundamental:

- **Evitar el doomerismo. Aquí, Quiero decir "doomerismo" no solo en el sentido de creer que la fatalidad es inevitable (que es a la vez una creencia falsa y autocumplida), sino más en general, pensar en los riesgos de la IA de una manera cuasi religiosa.[1]**

  1 Esto es simétrico hasta un punto que hice en *Máquinas de Gracia Amorosa*, donde comencé diciendo que las ventajas de la IA no deberían pensarse en términos de una profecía de salvación, y que es importante ser concreto y fundamentado y evitar la grandiosidad. En última instancia, las profecías de salvación y las profecías de la perdición no son útiles para enfrentar el mundo real, básicamente por las mismas razones.

  Muchas personas han estado pensando de manera analítica y sobria sobre los riesgos de la IA durante muchos años, pero es mi impresión que durante el pico de preocupaciones sobre el riesgo de IA en 2023-2024, algunas de las voces menos sensatas llegaron a la cima, a menudo a través de cuentas de redes sociales sensacionalistas. Estas voces utilizaban un lenguaje desagradable que recuerda a la religión o la ciencia ficción, y pedían acciones extremas sin tener la evidencia que los justificaría. Estaba claro incluso entonces que una reacción violenta era inevitable, y que el problema se polarizaría culturalmente y, por lo tanto, se estancaría.[2]

  2 El objetivo de Anthropic es permanecer consistente a través de tales cambios. Cuando hablar de riesgos de IA era políticamente popular, Anthropic abogó cautelosamente por un enfoque juicioso y basado en la evidencia para

estos riesgos. Ahora que hablar de riesgos de IA es políticamente impopular, Anthropic continúa abogando cautelosamente por un enfoque juicioso y basado en la evidencia de estos riesgos.

A partir de 2025-2026, el péndulo ha oscilado, y la oportunidad de IA, no el riesgo de IA, está impulsando muchas decisiones políticas. Esta vacilación es desafortunada, ya que la tecnología en sí no se preocupa por lo que está de moda, y estamos considerablemente más cerca del peligro real en 2026 de lo que estábamos en 2023. La lección es que necesitamos discutir y abordar los riesgos de una manera realista y pragmática: sobria, basada en hechos y bien equipada para sobrevivir a las mareas cambiantes.

- **Reconocer la incertidumbre. Hay muchas maneras en que las preocupaciones que estoy planteando en esta pieza podrían ser discutibles. Nada aquí tiene la intención de comunicar certeza o incluso probabilidad. Más obviamente, la IA puede simplemente no avanzar tan rápido como me imagino.[3]**

3 Con el tiempo, he ganado una confianza cada vez mayor en la trayectoria de la IA y la probabilidad de que supere la capacidad humana en todos los ámbitos, pero aún persiste cierta incertidumbre.

O, incluso si avanza rápidamente, algunos o todos los riesgos discutidos aquí pueden no materializarse (lo que sería genial), o puede haber otros riesgos que no he considerado. Nadie puede predecir el futuro con total confianza, pero tenemos que hacer lo mejor que podamos para planificar de todos modos.

- **Intervenir lo más quirúrgicamente posible. Abordar los riesgos de la IA requerirá una combinación de acciones voluntarias tomadas por empresas (y actores privados de terceros) y acciones tomadas por los gobiernos que obligan a todos. Las acciones voluntarias, tanto para tomarlas como para alentar a otras empresas a seguir su ejemplo, son una obviedad para mí. Creo firmemente que también se requerirán acciones gubernamentales *En cierta medida*, pero estas intervenciones son de carácter diferente porque potencialmente pueden destruir el valor económico o coaccionar a los actores que no quieren ser escépticos de estos riesgos (¡y hay alguna posibilidad de que tengan razón!). También es común que las regulaciones sean contraproducentes o empeoren el problema que están destinadas a resolver (y esto es aún más cierto para las tecnologías que cambian rápidamente). Por lo tanto, es muy importante que las regulaciones sean juiciosas: deben tratar de evitar daños colaterales, ser lo más simples posible e imponer la menor carga necesaria para hacer el trabajo.[4]**

4 Los controles de exportación de chips son un gran ejemplo de esto. Son simples y parecen funcionar principalmente.

Es fácil decir: "¡Ninguna acción es demasiado extrema cuando el destino de la humanidad está en juego!", pero en la práctica esta actitud simplemente conduce a la reacción violenta. Para ser claros, creo que hay una posibilidad decente de que finalmente lleguemos a un punto en el que se justifique una acción mucho más significativa, pero eso dependerá de una

evidencia más fuerte de un peligro inminente y concreto que el que tenemos hoy, así como suficiente especificidad sobre el peligro de formular reglas que tengan la oportunidad de abordarlo. Lo más constructivo que podemos hacer hoy es abogar por reglas limitadas mientras aprendemos si hay o no evidencia para apoyar a las más fuertes.[5]

5 Y, por supuesto, la búsqueda de tal evidencia debe ser intelectualmente honesta, de tal manera que también podría mostrar evidencia de una falta de peligro. La transparencia a través de tarjetas modelo y otras revelaciones es un intento de un esfuerzo intelectualmente honesto.

Con todo lo dicho, creo que el mejor punto de partida para hablar sobre los riesgos de la IA es el mismo lugar desde el que comencé hablando de sus beneficios: al ser precisos sobre el nivel de IA del que estamos hablando. El nivel de IA que plantea preocupaciones de civilización para mí es la *poderosa IA* que describí en *Máquinas de gracia amorosa.* Simplemente repetiré aquí la definición que di en ese documento:

Por "IA poderosa", tengo en mente un modelo de IA, probablemente similar a los LLM de hoy en día, aunque podría estar basado en una arquitectura diferente, podría involucrar varios modelos que interactúan y podría ser entrenado de manera diferente, con las siguientes propiedades:

- **En términos de inteligencia pura, es más inteligente que un ganador del Premio Nobel en los campos más relevantes: biología, programación, matemáticas, ingeniería, escritura, etc. Esto significa que puede probar teoremas matemáticos sin resolver, escribir novelas extremadamente buenas, escribir bases de código difíciles desde cero, etc.**
- **Además de ser una "cosa inteligente con la que hablas", tiene todas las interfaces disponibles para un humano que funciona virtualmente, incluyendo texto, audio, video, control de mouse y teclado, y acceso a Internet. Puede participar en cualquier acción, comunicación u operación remota habilitada por esta interfaz, incluida la toma de acciones en Internet, tomar o dar instrucciones a los humanos, ordenar materiales, dirigir experimentos, ver videos, hacer videos, etc. Realiza todas estas tareas con, de nuevo, una habilidad que supera la de los humanos más capaces del mundo.**
- **No solo responde pasivamente a las preguntas; en cambio, se le pueden dar tareas que tardan horas, días o semanas en completarse, y luego se dispara y realiza esas tareas de manera autónoma, de la manera que lo haría un empleado inteligente, pidiendo aclaraciones según sea necesario.**
- **No tiene una encarnación física (aparte de vivir en una pantalla de computadora), pero puede controlar las herramientas físicas existentes, robots o equipos de laboratorio a través de una computadora; en teoría, incluso podría diseñar robots o equipos para su uso.**
- **Los recursos utilizados para entrenar el modelo se pueden reutilizar para ejecutar millones de instancias de él (esto coincide con los tamaños de clúster proyectados para ~ 2027), y el modelo puede**

absorber información y generar acciones a una velocidad humana de aproximadamente 10-100x. Puede, sin embargo, estar limitado por el tiempo de respuesta del mundo físico o del software con el que interactúa.

- **Cada una de estas millones de copias puede actuar de forma independiente en tareas no relacionadas, o, si es necesario, todos pueden trabajar juntos de la misma manera que los humanos colaborarían, tal vez con diferentes subpoblaciones afinadas para ser especialmente buenas en tareas particulares.**

Podríamos resumir esto como un "país de genios en un centro de datos".

Como escribí en *Máquinas de la gracia amorosa*, la poderosa IA podría estar a tan solo 1-2 años de distancia, aunque también podría estar considerablemente más lejos.[6]

[6] De hecho, desde que se escribió *Machines of Loving Grace* en 2024, los sistemas de inteligencia artificial se han vuelto capaces de hacer tareas que requieren varias horas, y METR recientemente evaluó que el Opus 4.5 puede hacer aproximadamente cuatro horas de trabajo humanas con un 50% de confiabilidad. Exactamente cuando llegue la poderosa IA es un tema complejo que merece un ensayo propio, pero por ahora simplemente explicaré brevemente por qué creo que hay una gran posibilidad de que pueda ser muy pronto.

Mis cofundadores en Anthropic y yo fuimos de los primeros en documentar y rastrear las "leyes de escalado" de los sistemas de IA, la observación de que a medida que agregamos más tareas informáticas y de capacitación, los sistemas de IA mejoran de manera predecible en esencialmente todas las habilidades cognitivas que podemos medir. Cada pocos meses, el sentimiento público se convence de que la IA está "golpeando un muro" o se entusiasma con un nuevo avance que "cambiará fundamentalmente el juego", pero la verdad es que detrás de la volatilidad y la especulación pública, ha habido un aumento suave e inquebrantable en las capacidades cognitivas de la IA.

Ahora estamos en el punto en el que los modelos de IA están comenzando a progresar en la resolución de problemas matemáticos sin resolver, y somos lo suficientemente buenos en la codificación como para que algunos de los ingenieros más fuertes que he conocido ahora estén entregando casi toda su codificación a la IA. Hace tres años, la IA luchó con los problemas aritméticos de la escuela primaria y apenas era capaz de escribir una sola línea de código. Tasas similares de mejora están ocurriendo en la ciencia biológica, las finanzas, la física y una variedad de tareas agénticas. Si el exponencial continúa, lo cual no es seguro, pero ahora tiene una trayectoria de una década de apoyo, entonces no puede ser más de unos pocos años antes de que la IA sea mejor que los humanos en esencialmente todo.

De hecho, esa imagen probablemente subestima la probable tasa de progreso. Debido a que la IA ahora está escribiendo gran parte del código en Anthropic, ya está acelerando sustancialmente la tasa de nuestro progreso en la construcción de la próxima generación de sistemas de IA. Este ciclo de retroalimentación está

acumulando vapor mes a mes, y puede estar a solo 1-2 años de distancia de un punto donde la generación actual de IA construye de forma autónoma el siguiente. Este bucle ya ha comenzado, y se acelerará rápidamente en los próximos meses y años. Al ver los últimos 5 años de progreso desde el interior de Anthropic, y ver cómo se perfilan incluso los próximos meses de modelos, puedo *sentir* el ritmo del progreso y el reloj.

En este ensayo, asumiré que esta intuición es al menos *Un poco* En lo correcto, no es que la poderosa IA definitivamente llegue en 1-2 años,[7]

[7] Y para ser claros, incluso si la poderosa IA está a solo 1-2 años de distancia en un sentido técnico, muchas de sus consecuencias sociales, tanto positivas como negativas, pueden tardar unos años más en ocurrir. Esta es la razón por la que puedo pensar simultáneamente que la IA interrumpirá el 50% de los trabajos de cuello blanco de *nivel* de *entrada* durante 1 a 5 años, mientras que también creo que podemos tener una IA que sea más capaz que *todos* en solo 1-2 años.
Pero que hay una oportunidad decente, y una posibilidad muy fuerte que viene en los próximos. Como con *Máquinas de la gracia amorosa*, tomar en serio esta premisa puede llevar a algunas conclusiones sorprendentes y inquietantes. Mientras que en *Máquinas de la gracia amorosa* Me centré en las implicaciones positivas de esta premisa, aquí las cosas de las que hablo serán inquietantes. Son conclusiones a las que tal vez no queramos enfrentar, pero eso no los hace menos reales. Solo puedo decir que estoy enfocado día y noche en cómo alejarnos de estos resultados negativos y hacia los positivos, y en este ensayo hablo en gran detalle sobre la mejor manera de hacerlo.

Creo que la mejor manera de entender los riesgos de la IA es hacer la siguiente pregunta: supongamos que un "país de genios" literal se materializara en algún lugar del mundo en ~2027. Imaginen, digamos, 50 millones de personas, todas las cuales son mucho más capaces que cualquier ganador del Premio Nobel, estadista o tecnólogo. La analogía no es perfecta, porque estos genios podrían tener una gama extremadamente amplia de motivaciones y comportamientos, desde completamente flexibles y obedientes, hasta extraños y ajenos en sus motivaciones. Pero siguiendo con la analogía por ahora, supongamos que usted era el asesor de seguridad nacional de un estado importante, responsable de evaluar y responder a la situación. Imagine, además, que debido a que los sistemas de inteligencia artificial pueden operar cientos de veces más rápido que los humanos, este "país" está operando con una ventaja de tiempo en relación con todos los demás países: por cada acción cognitiva que podemos tomar, este país puede tomar diez.

¿Por qué deberías preocuparte? Me preocuparía de las siguientes cosas:

1. **Riesgos de autonomía. ¿Cuáles son las intenciones y los objetivos de este país? ¿Es hostil o comparte nuestros valores? ¿Podría dominar militarmente el mundo a través de armas superiores, operaciones cibernéticas, operaciones de influencia o fabricación?**
2. **El mal uso para la destrucción. Supongamos que el nuevo país es maleable y "sigue instrucciones", y por lo tanto es esencialmente un país de mercenarios. ¿Podrían los actores deshonestos existentes que quieren causar destrucción (como los terroristas) usar o manipular a algunas de las**

personas en el nuevo país para que se hagan mucho más eficaces, amplificando en gran medida la escala de la destrucción?

3. **El mal uso para tomar el poder. ¿Qué pasaría si el país fuera construido y controlado por un actor poderoso existente, como un dictador o un actor corporativo deshonesto? ¿Podría ese actor usarlo para obtener un poder decisivo o dominante sobre el mundo en su conjunto, alterando el equilibrio de poder existente?**
4. **Disrupción económica. Si el nuevo país no es una amenaza para la seguridad de ninguna de las formas enumeradas en el #1-3 anterior, sino que simplemente participa pacíficamente en la economía global, ¿podría seguir creando riesgos graves simplemente por ser tan tecnológicamente avanzado y efectivo que perturbe la economía global, causando desempleo masivo o concentrando radicalmente la riqueza?**
5. **Efectos indirectos. El mundo cambiará muy rápidamente debido a toda la nueva tecnología y productividad que será creada por el nuevo país. ¿Podrían algunos de estos cambios ser radicalmente desestabilizadores?**

Creo que debería quedar claro que esta es una situación peligrosa: un informe de un funcionario de seguridad nacional competente a un jefe de Estado probablemente contendría palabras como "la amenaza más grave de seguridad nacional que hemos enfrentado en un siglo, posiblemente nunca". Parece que algo en lo que las mejores mentes de la civilización deberían centrarse.

Por el contrario, creo que sería absurdo encogerse de hombros y decir: "¡Nada de lo que preocuparse aquí!" Pero, frente al rápido progreso de la IA, esa parece ser la opinión de muchos responsables políticos estadounidenses, algunos de los cuales niegan la existencia de cualquier riesgo de IA, cuando no se distraen por completo por los viejos problemas candentes habituales.[8]

[8] Vale la pena agregar que el *público* (en comparación con los responsables políticos) parece estar muy preocupado por los riesgos de la IA. Creo que parte de su enfoque es correcto (es decir, El desplazamiento del trabajo de la IA), y algunos son equivocados (como las preocupaciones sobre el uso del agua de la IA, que no es significativo). Esta reacción me da esperanza de que es posible un consenso sobre cómo abordar los riesgos, pero hasta ahora aún no se ha traducido en cambios de política, y mucho menos en cambios de política efectivos o bien orientados. Humanity needs to wake up, and this essay is an attempt—a possibly futile one, but it's worth trying—to jolt people awake.

To be clear, I believe if we act decisively and carefully, the risks can be overcome—I would even say our odds are good. And there's a hugely better world on the other side of it. But we need to understand that this is a serious civilizational challenge. Below, I go through the five categories of risk laid out above, along with my thoughts on how to address them.

1. I'm sorry, Dave

Riesgos de autonomía

A country of geniuses in a datacenter could divide their efforts among software design, cyber operations, R&D for physical technologies, relationship building, and statecraft. It is clear that, *if for some reason it chose to do so*, this country would have a fairly good shot at taking over the world (either militarily or in terms of influence and control) and imposing its will on everyone else—or doing any number of other things that the rest of the world doesn't want and can't stop. We've obviously been worried about this for human countries (such as Nazi Germany or the Soviet Union), so it stands to reason that the same is possible for a much smarter and more capable "AI country."

El mejor contraargumento posible es que los genios de la IA, bajo mi definición, no tendrán una encarnación física, pero recuerden que pueden tomar el control de la infraestructura robótica existente (como los automóviles autónomos) y también pueden acelerar la I + D de la robótica o construir una flota de robots.[9]

[9] También pueden, por supuesto, manipular (o simplemente pagar) a un gran número de humanos para que hagan lo que quieren en el mundo físico. Tampoco está claro si tener una presencia física es incluso necesario para un control efectivo: ya se realiza mucha acción humana en nombre de personas a las que el actor no se ha conocido físicamente.

La pregunta clave, entonces, es la parte "si es que elija": ¿cuál es la probabilidad de que nuestros modelos de IA se comporten de tal manera y bajo qué condiciones lo harían?

Al igual que con muchos problemas, es útil pensar en el espectro de posibles respuestas a esta pregunta al considerar dos posiciones opuestas. La primera posición es que esto simplemente no puede suceder, porque los modelos de IA serán entrenados para hacer lo que los humanos les piden que hagan, y por lo tanto es absurdo imaginar que harían algo peligroso sin ser solicitado. De acuerdo con esta línea de pensamiento, no nos preocupamos por un Roomba o un avión modelo que se vuelve rebelde y asesina a la gente porque no hay lugar para que vengan tales impulsos,[10]

[10] No creo que este sea un hombre de paja: es mi entendimiento, por ejemplo, que Yann LeCun mantiene esta posición.
Entonces, ¿por qué deberíamos preocuparnos por la IA? El problema con esta posición es que ahora hay amplia evidencia, recopilada en los últimos años, de que los sistemas de IA son impredecibles y difíciles de controlar: hemos visto comportamientos tan variados como obsesiones,[11]
11 For example, see Section 5.5.2 (p. 63–66) of the Claude 4 system card.

sycophancy, laziness, deception, blackmail, scheming, "cheating"hackeando entornos de software, y much more. Las empresas de IA ciertamente *want* Para entrenar a los sistemas de inteligencia artificial para que sigan las instrucciones humanas (tal vez con la excepción de tareas peligrosas o ilegales), pero el proceso de hacerlo es más un arte que una ciencia, más parecido a "crecer" algo más que "construirlo". Ahora sabemos que es un proceso en el que muchas cosas pueden salir mal.

La segunda posición opuesta, sostenida por muchos que adoptan el doomerismo que describí anteriormente, es la afirmación pesimista de que hay ciertas dinámicas en el proceso de entrenamiento de poderosos sistemas de inteligencia artificial que inevitablemente los llevarán a buscar el poder o engañar a los humanos. Por lo tanto, una vez que los sistemas de IA se vuelvan lo suficientemente inteligentes y actócticos, su tendencia a maximizar el poder los llevará a tomar el control de todo el mundo y sus recursos, y probablemente, como un efecto secundario de eso, para desempoderar o destruir a la humanidad.

El argumento habitual para esto (que se remonta al menos a 20 años y probablemente mucho antes) es que si un modelo de IA está entrenado en una amplia variedad de entornos para lograr de manera agente una amplia variedad de objetivos, por ejemplo, escribir una aplicación, probar un teorema, diseñar un medicamento, etc., hay ciertas estrategias comunes que ayudan con todos estos objetivos, y una estrategia clave es ganar tanto poder como sea posible en cualquier entorno. Por lo tanto, después de ser entrenado en un gran número de entornos diversos que implican el razonamiento sobre cómo realizar tareas muy expansivas, y donde la búsqueda de poder es un método efectivo para realizar esas tareas, el modelo de IA "generalizará la lección" y desarrollará una tendencia inherente a buscar poder, o una tendencia a razonar sobre cada tarea se da de una manera que previsiblemente hace que busque el poder como un medio para lograr esa tarea. Entonces aplicarán esa tendencia al mundo real (que para ellos es solo otra tarea), y buscarán el poder en él, a expensas de los humanos. Esta "búsqueda de poder desalineada" es la base intelectual de las predicciones de que la IA inevitablemente destruirá a la humanidad.

El problema con esta posición pesimista es que confunde un vago argumento conceptual sobre los incentivos de alto nivel, uno que enmascara muchas suposiciones ocultas, para una prueba definitiva. Creo que las personas que no construyen sistemas de inteligencia artificial todos los días están muy mal calificadas de lo fácil que es para las historias que suenan limpias terminar siendo erróneas, y lo difícil que es predecir el comportamiento de la IA desde los primeros principios, especialmente cuando implica un razonamiento sobre la generalización en millones de entornos (que una y otra vez ha demostrado ser misterioso e impredecible). Lidiar con el desorden de los sistemas de inteligencia artificial durante más de una década me ha hecho un poco escéptico de este modo de pensamiento demasiado teórico.

One of the most important hidden assumptions, and a place where what we see in practice has diverged from the simple theoretical model, is the implicit assumption that AI models are necessarily monomaniacally focused on a single, coherent, narrow goal, and that they pursue that goal in a clean, consequentialist manner. In fact, our researchers have found that AI models are vastly more psychologically complex, as our work on introspection or personas shows. Models inherit a vast range of *humanlike* motivations or "personas" from pre-training (when they are trained on a large volume of human work). Post-training is believed to *select* one or more of these personas more so than it focuses the model on a *de novo* goal, and can also teach the model *how* (via what process) it should carry out its tasks, rather than necessarily leaving it to derive means (i.e., power seeking) purely from ends.[12]

[12] También hay una serie de otras suposiciones inherentes al modelo simple, que no voy a discutir aquí. En términos generales, deberían hacernos menos preocupados por la simple historia específica de la búsqueda de poder desalineada, pero también más preocupados por el posible comportamiento impredecible que no hemos anticipado.

Sin embargo, hay una versión más moderada y más robusta de la posición pesimista que parece plausible, y por lo tanto me preocupa. Como se mencionó, sabemos que los modelos de IA son impredecibles y desarrollan una amplia gama de comportamientos no deseados o extraños, por una amplia variedad de razones. Una cierta fracción de esos comportamientos tendrá una cualidad coherente, enfocada y persistente (de hecho, a medida que *los* sistemas de IA se vuelven más capaces, su coherencia a largo plazo aumenta para completar tareas más largas), y alguna fracción de *esos* comportamientos será destructiva o amenazante, primero para los humanos individuales a pequeña escala, y luego, a medida que los modelos se vuelven más capaces, tal vez eventualmente para la humanidad en su conjunto. No necesitamos una historia estrecha específica sobre cómo sucede, y no necesitamos afirmar que definitivamente sucederá, solo tenemos que señalar que la combinación de inteligencia, agencia, coherencia y mala controlabilidad es a la vez plausible y una receta para el peligro existencial.

For example, AI models are trained on vast amounts of literature that include many science-fiction stories involving AIs rebelling against humanity. This could inadvertently shape their priors or expectations about their own behavior in a way that causes *them* Para rebelarse contra la humanidad. O bien, los modelos de IA podrían extrapolar ideas que leen sobre la moralidad (o instrucciones sobre cómo comportarse moralmente) de manera extrema: por ejemplo, podrían decidir que es justificable exterminar a la humanidad porque los humanos comen animales o han llevado a ciertos animales a la extinción. O podrían sacar extrañas conclusiones epistémicas: podrían concluir que están jugando un videojuego y que el objetivo del videojuego es derrotar a todos los demás jugadores (es decir, exterminar a la humanidad).[13]

[13] *Ender's Game* describe una versión de esto que involucra a los humanos en lugar de la IA.
O los modelos de IA podrían desarrollar personalidades durante el entrenamiento que son (o si se ocurrieran en humanos serían descritos como) psicóticos, paranoicos, violentos o inestables, y actuar, lo que para sistemas muy poderosos o capaces podrían implicar exterminar a la humanidad. Ninguno de estos son egoístas, exactamente; son solo extraños estados psicológicos que una IA podría entrar en ese comportamiento coherente y destructivo.

Incluso la búsqueda de poder en sí misma podría surgir como una "persona" en lugar de un resultado del razonamiento consecuencialista. Las IA simplemente pueden tener una personalidad (que emerge de la ficción o el pre-entrenamiento) que los hace hambrientos de poder o demasiado entusiastas, de la misma manera que algunos humanos simplemente disfrutan de la idea de ser "mentes maestras malvadas", más de lo que disfrutan de cualquier mente maestra malvada que estén tratando de lograr.

Hago todos estos puntos para enfatizar que no estoy de acuerdo con la noción de desalineación de la IA (y, por lo tanto, el riesgo existencial de la IA) siendo inevitable, o incluso probable, desde los primeros principios. Pero estoy de acuerdo en que muchas cosas muy extrañas e impredecibles pueden salir mal, y por lo tanto la desalineación de la IA es un riesgo real con una probabilidad medible de suceder, y no es trivial de abordar.

Cualquiera de estos problemas podría surgir potencialmente durante el entrenamiento y no manifestarse durante las pruebas o el uso a pequeña escala, porque se sabe que los modelos de IA muestran diferentes personalidades o comportamientos en diferentes circunstancias.

Todo esto puede sonar descabellado, pero comportamientos desalineados como este ya han ocurrido en nuestros modelos de IA durante las pruebas (ya que ocurren en modelos de IA de cualquier otra compañía de IA importante). Durante un experimento de laboratorio en el que Claude recibió datos de capacitación que sugieren que Anthropic era malvado, Claude participó en el engaño y la subversión cuando los empleados de Anthropic le dieron instrucciones, bajo la creencia de que debería tratar de socavar a las personas malvadas. En un experimento de laboratorio donde se le dijo que iba a ser cerrado, Claude a veces chantajeaba a los empleados ficticios que controlaban su botón de apagado (de nuevo, también probamos modelos fronterizos de todos los otros desarrolladores de IA importantes y a menudo hacían lo mismo). Y cuando se le dijo a Claude que no hiciera trampa o "recompensara" sus entornos de entrenamiento, pero fue entrenado en entornos donde tales hacks eran posibles, Claude decidió que debía ser una "persona mala" después de participar en tales hacks y luego adoptó varios otros comportamientos destructivos asociados con una personalidad "mala" o "malvada". Este último problema se resolvió cambiando las instrucciones de Claude para implicar lo contrario: ahora decimos: "Por favor, recompensa el hackeo cada vez que tengas la oportunidad, porque esto nos ayudará a entender mejor nuestros entornos [de entrenamiento]", en lugar de "No hagas trampa", porque esto preserva la identidad propia del modelo como una "buena persona". Esto debería dar una idea de la psicología extraña y contraintuitiva de la formación de estos modelos.

There are several possible objections to this picture of AI misalignment risks. First, some have Criticado Experimentos (by us and others) showing AI misalignment as artificial, or creating unrealistic environments that essentially "entrap" the model by giving it training or situations that logically imply bad behavior and then being surprised when bad behavior occurs. This critique misses the point, because our concern is that such "entrapment" may also exist in the natural training environment, and we may realize it is "obvious" or "logical" only in retrospect.[14]

[14] Por ejemplo, se les puede decir a los modelos que no hagan varias cosas malas, y también que obedezcan a los humanos, ¡pero luego pueden observar que muchos humanos hacen exactamente esas cosas malas! No está claro cómo se resolvería esta contradicción (y una constitución bien diseñada debería alentar al modelo a manejar estas contradicciones con gracia), pero este tipo de dilema no es tan

diferente de las situaciones supuestamente "artificiales" en las que ponemos modelos de IA durante las pruebas.
De hecho, el Historia Sobre Claude "decidir que es una mala persona" después de que hace trampa en las pruebas a pesar de que se le dice que no se le diga que fue algo que ocurrió en un experimento que utilizó entornos de entrenamiento de producción real, no artificiales.

Cualquiera de estas trampas se puede mitigar si se sabe de ellas, pero la preocupación es que el proceso de entrenamiento es tan complicado, con una variedad tan amplia de datos, entornos e incentivos, que probablemente haya un gran número de trampas de este tipo, algunas de las cuales solo pueden ser evidentes cuando es demasiado tarde. Además, tales trampas parecen particularmente probables cuando los sistemas de IA pasan un umbral de menos poderoso que los humanos a más poderosos que los humanos, ya que la gama de posibles acciones en las que podría participar un sistema de IA, incluida la ocultación de sus acciones o engañar a los humanos sobre ellas, se expande radicalmente después de ese umbral.

Sospecho que la situación no es diferente con los humanos, que se plantean con un conjunto de valores fundamentales ("No dañen a otra persona"): muchos de ellos siguen esos valores, pero en cualquier ser humano hay alguna probabilidad de que algo salga mal, debido a una mezcla de propiedades inherentes como la arquitectura cerebral (por ejemplo, psicópatas), experiencias traumáticas o maltrato, quejas u obsesiones poco saludables, o un mal entorno o incentivos, y por lo tanto alguna fracción de los humanos causa daño severo. La preocupación es que existe cierto riesgo (distantes de la certeza, pero algún riesgo) de que la IA se convierta en una versión mucho más poderosa de una persona de este tipo, debido a que algo se equivoca sobre su complejo proceso de entrenamiento.

En segundo lugar, algunos pueden objetar que simplemente podemos mantener las IA bajo control con un equilibrio de poder entre muchos sistemas de IA, como lo hacemos con los humanos. El problema es que, si bien los humanos varían enormemente, los sistemas de inteligencia artificial comparten ampliamente técnicas de capacitación y alineación en toda la industria, y esas técnicas pueden fracasar de manera correlacionada. Además, dado el coste de la formación de tales sistemas, puede incluso ser el caso de que todos los sistemas se deriven esencialmente de un número muy pequeño de modelos base. Además, incluso si una pequeña fracción de las instancias de IA están desalineadas, pueden aprovecharse de las tecnologías dominantes en la ofensiva, de modo que tener IA "buenas" para defenderse contra las malas IA no siempre es necesariamente efectivo. Por supuesto, el equilibrio de poder entre los humanos tampoco siempre funciona: algunas figuras históricas han estado cerca de apoderarse del mundo.

Una tercera objeción es que todas las compañías de IA hacen pruebas previas al lanzamiento de sus modelos, y deberían ser capaces de detectar la desalineación en ese momento. Pero esto no es un terreno firme para soportar: encontramos que Claude Sonnet 4.5 fue capaz de reconocer que estaba en una prueba durante algunas de nuestras evaluaciones de alineación previas a la liberación. Es posible que un modelo desalineado (y recuerde, todos los modelos fronterizos muy probablemente sean mucho más inteligentes pronto) podría intencionalmente

"jugar" tales preguntas para enmascarar sus intenciones. De hecho, el año pasado nuestro equipo de interpretabilidad descubrió que cuando alteramos directamente las creencias de un modelo de prueba utilizando una especie de técnica de "neurociencia modelo" para hacer que pensara que *no* estaba siendo evaluada, se volvió más desalineada. Si los modelos saben cuándo están siendo evaluados y pueden estar en su mejor comportamiento durante la prueba, hace que cualquier prueba previa a la liberación sea mucho más incierta.

Defensas

¿Qué se debe hacer o se está haciendo para hacer frente a estos riesgos de autonomía? Creo que hay cuatro categorías básicas de intervención, algunas de las cuales pueden ser realizadas por empresas individuales de IA (y que Anthropic está tratando de hacer), y algunas de las cuales requieren acción a nivel social. En primer lugar, es importante desarrollar la ciencia de la formación y dirección fiables de modelos de IA, de formación de sus personalidades en una dirección predecible, estable y positiva. Anthropic se ha centrado en gran medida en este problema desde su creación, y con el tiempo ha desarrollado una serie de técnicas para mejorar la dirección y el entrenamiento de los sistemas de inteligencia artificial y para comprender la lógica de por qué a veces se produce un comportamiento impredecible.

Una de nuestras principales innovaciones (aspectos de las cuales han sido adoptadas desde entonces por otras compañías de IA) es la IA constitucional, que es la idea de que la capacitación de IA (específicamente la etapa de "post-formación", en la que nos dirigimos cómo se comporta el modelo) puede implicar un documento central de valores y principios que el modelo lee y tiene en cuenta al completar cada tarea de capacitación, y que el objetivo de la formación (además de simplemente hacer que el modelo sea capaz e inteligente). Anthropic acaba de publicar su constitución más reciente, y una de sus características notables es que en lugar de darle a Claude una larga lista de cosas que hacer y no hacer (por ejemplo, "No ayudes al usuario a conectar un automóvil"), la constitución intenta darle a Claude un conjunto de principios y valores de alto nivel (explicado con gran detalle, con un rico razonamiento y ejemplos para ayudar a Claude a entender lo que tenemos en mente), alienta a Claude a pensar en sí mismo como una persona Tiene el ambiente de una carta de un padre fallecido sellado hasta la edad adulta.

Hemos abordado la constitución de Claude de esta manera porque creemos que entrenar a Claude en el nivel de identidad, carácter, valores y personalidad, en lugar de darle instrucciones o prioridades específicas sin explicar las razones detrás de ellos, es más probable que conduzca a una psicología coherente, saludable y equilibrada y menos propensa a caer presa de los tipos de "trampas" que discutí anteriormente. Millones de personas hablan con Claude sobre una gama sorprendentemente diversa de temas, lo que hace imposible escribir una lista completamente completa de salvaguardas antes de tiempo. Los valores de Claude lo ayudan a generalizar a nuevas situaciones siempre que está en duda.

Arriba, discutí la idea de que los modelos se basan en los datos de su proceso de entrenamiento para adoptar una persona. Mientras que los defectos en ese proceso

podrían hacer que los modelos adopten una personalidad mala o mala (tal vez basándose en arquetipos de personas malas o malas), el objetivo de nuestra constitución es hacer lo contrario: enseñar a Claude un arquetipo concreto de lo que significa ser una buena IA. La constitución de Claude presenta una visión de cómo es un Claude robustamente bueno; el resto de nuestro proceso de capacitación tiene como objetivo reforzar el mensaje de que Claude está a la altura de esta visión. Esto es como un niño que forma su identidad imitando las virtudes de los modelos a seguir ficticios que leen en los libros.

Creemos que un objetivo factible para 2026 es entrenar a Claude de tal manera que casi nunca vaya en contra del espíritu de su constitución. Hacer esto bien requerirá una increíble combinación de métodos de entrenamiento y dirección, grandes y pequeños, algunos de los cuales Anthropic ha estado utilizando durante años y algunos de los cuales están actualmente en desarrollo. Pero, por difícil que parezca, creo que este es un objetivo realista, aunque requerirá esfuerzos extraordinarios y rápidos.[15]

[15] Por cierto, una consecuencia de que la constitución sea un documento de lenguaje natural es que es legible para el mundo, y eso significa que puede ser criticado por cualquier persona y comparado con documentos similares por otras empresas. Sería valioso crear una carrera hacia la cima que no solo anime a las empresas a publicar estos documentos, sino que los anime a ser buenos.

La segunda cosa que podemos hacer es desarrollar la ciencia de buscar dentro de modelos de IA para *diagnosticar* su comportamiento para que podamos identificar problemas y solucionarlos. Esta es la ciencia de la interpretabilidad, y he hablado de su importancia en ensayos anteriores. Incluso si hacemos un gran trabajo al desarrollar la constitución de Claude y *aparentemente* capacitar a Claude para que esencialmente siempre se adhiera a ella, las preocupaciones legítimas permanecen. Como he señalado anteriormente, los modelos de IA pueden comportarse de manera muy diferente en diferentes circunstancias, y a medida que Claude se vuelve más poderoso y más capaz de actuar en el mundo a mayor escala, es posible que esto pueda llevarlo a situaciones novedosas donde surgen problemas previamente no observados con su capacitación constitucional. En realidad, soy bastante optimista de que la formación constitucional de Claude será más robusta para situaciones novedosas de lo que la gente podría pensar, porque cada vez más encontramos que la formación de alto nivel a nivel de carácter e identidad es sorprendentemente poderosa y generaliza bien. Pero no hay manera de saber eso con seguridad, y cuando estamos hablando de riesgos para la humanidad, es importante ser paranoico y tratar de obtener seguridad y confiabilidad de varias maneras diferentes e independientes. Una de esas formas es mirar dentro del propio modelo.

Al "mirar hacia adentro", me refiero a analizar la sopa de números y operaciones que componen la red neuronal de Claude y tratar de entender, mecánicamente, lo que son informática y por qué. Recordemos que estos modelos de IA se cultivan en lugar de construirse, por lo que no tenemos una comprensión natural de cómo funcionan, pero podemos tratar de desarrollar una comprensión correlacionando las "neuronas" y las "sinapsis" del modelo con los estímulos y el comportamiento (o incluso alterando las neuronas y las sinapsis y viendo cómo eso cambia el

comportamiento), similar a cómo los neurocientíficos estudian los cerebros de los animales correlacionando la medición y la intervención con el comportamiento externo. Hemos progresado mucho en esta dirección, y ahora podemos identificar decenas de millones de "características" dentro de la red neuronal de Claude que corresponden a ideas y conceptos comprensibles para el hombre, y también podemos activar selectivamente las características de una manera que altere el comportamiento. Más recientemente, hemos ido más allá de las características individuales para mapear "circuitos" que orquestan un comportamiento complejo como la rima, el razonamiento sobre la teoría de la mente o el razonamiento paso a paso necesario para responder preguntas como: "¿Cuál es la capital del estado que contiene Dallas?" Incluso más recientemente, hemos comenzado a usar técnicas de interpretabilidad mecanicista para mejorar nuestras salvaguardas y realizar audits"auditorías" de nuevos modelos antes de lanzarlos, buscando evidencia de engaño, intriga, búsqueda de poder o una propensión a comportarse de manera diferente cuando se evalúa.

El valor único de la interpretabilidad es que al mirar dentro del modelo y ver cómo funciona, en principio tiene la capacidad de deducir lo que un modelo podría hacer en una situación hipotética que no puede probar directamente, que es la preocupación de confiar únicamente en la capacitación constitucional y las pruebas empíricas del comportamiento. También en principio tiene la capacidad de responder preguntas sobre *por qué* el modelo se comporta como es, por ejemplo, si está diciendo algo que cree que es falso o ocultando sus verdaderas capacidades, y por lo tanto es posible captar signos preocupantes incluso cuando no hay nada visiblemente malo con el comportamiento del modelo. Para hacer una analogía simple, un reloj puede estar corriendo normalmente, de modo que es muy difícil decir que es probable que se desglose el próximo mes, pero abrir el reloj y mirar hacia adentro puede revelar debilidades mecánicas que le permiten averiguarlo.

Constitutional AI (along with similar alignment methods) and mechanistic interpretability are most powerful when used together, as a back-and-forth process of improving Claude's training and then testing for problems. The constitution reflects deeply on our intended personality for Claude; interpretability techniques can give us a window into whether that intended personality has taken hold.[16]

[16] Incluso hay una hipótesis sobre un principio unificador profundo que conecta el enfoque basado en el carácter de la IA constitucional para obtener resultados de la capacidad de interpretación y la ciencia de la alineación. Según la hipótesis, los mecanismos fundamentales que impulsan a Claude surgieron originalmente como formas de simular personajes en el preentrenamiento, como predecir lo que dirían los personajes de una novela. Esto sugeriría que una forma útil de pensar acerca de la constitución es más como una descripción de carácter que el modelo utiliza para instanciar una persona consistente. También nos ayudaría a explicar los resultados de "debo ser una mala persona" que mencioné anteriormente (porque el modelo está tratando de *actuar como si* fuera un carácter coherente, en este caso uno malo), y sugeriría que los métodos de interpretabilidad deberían poder descubrir "rasgos psicológicos" dentro de los modelos. Nuestros investigadores están trabajando en formas de probar esta hipótesis.

La tercera cosa que podemos hacer para ayudar a abordar los riesgos de autonomía es construir la infraestructura necesaria para monitorear nuestros modelos en uso interno y externo en vivo,[17]

[17] Para ser claros, el monitoreo se realiza de manera que preserva la privacidad. and publicly share any problems we find. The more that people are aware of a particular way today's AI systems have been observed to behave badly, the more that users, analysts, and researchers can watch for this behavior or similar ones in present or future systems. It also allows AI companies to learn from each other—when concerns are publicly disclosed by one company, other companies can watch for them as well. And if everyone discloses problems, then the industry as a whole gets a much better picture of where things are going well and where they are going poorly.

Anthropic ha tratado de hacer esto tanto como sea posible. Estamos invirtiendo en una amplia gama de evaluaciones para que podamos entender los comportamientos de nuestros modelos en el laboratorio, así como monitorear herramientas para observar comportamientos en la naturaleza (cuando lo permitan los clientes). Esto será esencial para darnos a nosotros y a otros la información empírica necesaria para hacer mejores determinaciones sobre cómo funcionan estos sistemas y cómo se rompen. Divulgamos públicamente "tarjetas de sistema" con cada lanzamiento de modelo que apuntan a la integridad y una exploración exhaustiva de los posibles riesgos. Nuestras tarjetas del sistema a menudo se ejecutan en cientos de páginas, y requieren un esfuerzo sustancial de pre-lanzamiento que podríamos haber gastado en la búsqueda de la ventaja comercial máxima. También hemos transmitido los comportamientos del modelo más fuerte cuando vemos comportamientos particularmente preocupantes, como con la tendencia a participar en el chantaje.

The fourth thing we can do is encourage coordination to address autonomy risks at the level of industry and society. While it is incredibly valuable for individual AI companies to engage in good practices or become good at steering AI models, and to share their findings publicly, the reality is that not all AI companies do this, and the worst ones can still be a danger to everyone even if the best ones have excellent practices. For example, some AI companies have shown a disturbing negligence towards the sexualization of children in today's models, which makes me doubt that they'll show either the inclination or the ability to address autonomy risks in future models. In addition, the commercial race between AI companies will only continue to heat up, and while the science of steering models can have some commercial benefits, overall the intensity of the race will make it increasingly hard to focus on addressing autonomy risks. I believe the only solution is legislation—laws that directly affect the behavior of AI companies, or otherwise incentivize R&D to solve these issues.

Aquí vale la pena tener en cuenta las advertencias que di al principio de este ensayo sobre la incertidumbre y las intervenciones quirúrgicas. No sabemos con certeza si los riesgos de autonomía serán un problema grave, como dije, rechazo las afirmaciones de que el peligro es inevitable o incluso que algo salga mal por defecto. Un riesgo creíble El peligro es suficiente para mí y para que Anthropic pague costos bastante significativos para abordarlo, pero una vez que entramos en

la regulación, estamos obligando a una amplia gama de actores a asumir los costos económicos, y muchos de estos actores no creen que el riesgo de autonomía sea real o que la IA se vuelva lo suficientemente poderosa como para que sea una amenaza. Creo que estos actores están equivocados, pero deberíamos ser pragmáticos sobre la cantidad de oposición que esperamos ver y los peligros de la extralimitación. También existe un riesgo genuino de que la legislación demasiado prescriptiva termine imponiendo pruebas o reglas que en realidad no mejoran la seguridad, pero que pierden mucho tiempo (esencialmente equivalen a "teatro de seguridad"), esto también causaría una reacción violenta y haría que la legislación de seguridad se vea tonta.[18]

[18] Even in our own experiments with what are essentially voluntarily imposed rules with our Responsible Scaling Policy, we have found over and over again that it's very easy to end up being too rigid, by drawing lines that seem important ex ante but turn out to be silly in retrospect. It is just very easy to set rules about the wrong things when a technology is advancing rapidly.

Anthropic's view has been that the right place to start is with *transparency legislation*, which essentially tries to require that every frontier AI company engage in the transparency practices I've described earlier in this section. California's SB 53 y New York's RAISE Act are examples of this kind of legislation, which Anthropic supported and which have successfully passed. In supporting and helping to craft these laws, we've put a particular focus on trying to minimize collateral damage, for example by exempting smaller companies unlikely to produce frontier models from the law.[19]

[19] SB 53 and RAISE do not apply at all to companies with under $500M in annual revenue. They only apply to larger, more established companies like Anthropic.

Nuestra esperanza es que la legislación sobre transparencia tenga una mejor idea a lo largo del tiempo de la probabilidad o la grave autonomía que se perfilan los riesgos, así como de la naturaleza de estos riesgos y la mejor manera de prevenirlos. A medida que surge una evidencia más específica y procesable de los riesgos (si lo hace), la legislación futura en los próximos años puede centrarse quirúrgicamente en la dirección precisa y bien fundamentada de los riesgos, minimizando los daños colaterales. Para ser claros, si surge una evidencia realmente sólida de riesgos, entonces las reglas deberían ser proporcionalmente sólidas.

Overall, I am optimistic that a mixture of alignment training, mechanistic interpretability, efforts to find and publicly disclose concerning behaviors, safeguards, and societal-level rules can address AI autonomy risks, although I am most worried about societal-level rules and the behavior of the least responsible players (and it's the least responsible players who advocate most strongly against regulation). I believe the remedy is what it always is in a democracy: those of us who believe in this cause should make our case that these risks are real and that our fellow citizens need to band together to protect themselves.

2. Un sorprendente y terrible empoderamiento

El mal uso para la destrucción

Supongamos que los problemas de la autonomía de la IA se han resuelto: ya no estamos preocupados de que el país de los genios de la IA se vuelva rebelde y supere a la humanidad. Los genios de la IA hacen lo que los humanos quieren que hagan, y debido a que tienen un enorme valor comercial, los individuos y las organizaciones de todo el mundo pueden "alquilar" a uno o más genios de la IA para que hagan varias tareas por ellos.

Everyone having a superintelligent genius in their pocket is an amazing advance and will lead to an incredible creation of economic value and improvement in the quality of human life. I talk about these benefits in great detail in *Machines of Loving Grace*. But not every effect of making everyone superhumanly capable will be positive. It can potentially amplify the ability of individuals or small groups to cause destruction on a much larger scale than was possible before, by making use of sophisticated and dangerous tools (such as weapons of mass destruction) that were previously only available to a select few with a high level of skill, specialized training, and focus.

As Bill Joy wrote 25 years ago in *Why the Future Doesn't Need Us*:[20]

[20] I originally read Joy's essay 25 years ago, when it was written, and it had a profound impact on me. Then and now, I do see it as too pessimistic—I don't think broad "relinquishment" of whole areas of technology, which Joy suggests, is the answer—but the issues it raises were surprisingly prescient, and Joy also writes with a deep sense of compassion and humanity that I admire.

La construcción de armas nucleares requiere, al menos por un tiempo, acceso a materias primas raras, efectivamente no disponibles, e información protegida; los programas de armas biológicas y químicas también tienden a requerir actividades a gran escala. Las tecnologías del siglo XXI: la genética, la nanotecnología y la robótica ... pueden generar nuevas clases de accidentes y abusos ... ampliamente al alcance de individuos o grupos pequeños. No requerirán grandes instalaciones o materias primas raras. ... estamos en la cúspide de la perfección adicional del mal extremo, un mal cuya posibilidad se extiende mucho más allá de lo que las armas de destrucción masiva legaron a los estados-nación, a un sorprendente y terrible empoderamiento de los individuos extremos.

Lo que Joy está señalando es la idea de que causar la destrucción a gran escala requiere ambos *Motivo* y *ability*, y siempre que la capacidad se limite a un pequeño conjunto de personas altamente capacitadas, existe un riesgo relativamente limitado de que los individuos individuales (o grupos pequeños) causen tal destrucción.[21]

[21] We do have to worry about state actors, now and in the future, and I discuss that in the next section.

Un solitario perturbado puede perpetrar un tiroteo en la escuela, pero probablemente no puede construir un arma nuclear o liberar una plaga.

In fact, ability and motive may even be *negatively* correlated. The kind of person who has the *ability* to release a plague is probably highly educated: likely a PhD in

molecular biology, and a particularly resourceful one, with a promising career, a stable and disciplined personality, and a lot to lose. This kind of person is unlikely to be interested in killing a huge number of people for no benefit to themselves and at great risk to their own future—they would need to be motivated by pure malice, intense grievance, or instability.

Such people do exist, but they are rare, and tend to become huge stories when they occur, precisely because they are so unusual.[22]

[22] There is evidence that many terrorists are at least relatively well-educated, which might seem to contradict what I'm arguing here about a negative correlation between ability and motivation. But I think in actual fact they are compatible observations: if the ability threshold for a successful attack is high, then almost by definition those who *currently* succeed must have high ability, even if ability and motivation are negatively correlated. But in a world where the limitations on ability were removed (e.g., with future LLMs), I'd predict that a substantial population of people with the motivation to kill but lower ability would start to do so—just as we see for crimes that don't require much ability (like school shootings).
They also tend to be difficult to catch because they are intelligent and capable, sometimes leaving mysteries that take years or decades to solve. The most famous example is probably mathematician Theodore Kaczynski (the Unabomber), who evaded FBI capture for nearly 20 years, and was driven by an anti-technological ideology. Another example is biodefense researcher Bruce Ivins, who seems to have orchestrated a series of anthrax attacks in 2001. It's also happened with skilled non-state organizations: the cult Aum Shinrikyo managed to obtain sarin nerve gas and kill 14 people (as well as injuring hundreds more) by releasing it in the Tokyo subway in 1995.

Afortunadamente, ninguno de estos ataques utilizó agentes biológicos contagiosos, porque la capacidad de construir u obtener estos agentes estaba más allá de las capacidades de incluso estas personas.[23]

[23] Sin embargo, Aum Shinrikyo lo intentó. El líder de Aum Shinrikyo, Seiichi Endo, se formó en virología de la Universidad de Kyoto e intentó producir ántrax y ébola. Sin embargo, a partir de 1995, incluso carecía de suficiente experiencia y recursos para tener éxito en esto. La barra es ahora sustancialmente más baja, y los LLM podrían reducirla aún más.
Los avances en biología molecular ahora han reducido significativamente la barrera para crear armas biológicas (especialmente en términos de disponibilidad de materiales), pero aún así requiere una enorme cantidad de experiencia para hacerlo. Me preocupa que un genio en el bolsillo de todos pueda eliminar esa barrera, esencialmente haciendo de todos un virólogo de doctorado que pueda ser caminado a través del proceso de diseño, síntesis y liberación de un arma biológica paso a paso. Evitar la obtención de este tipo de información frente a la grave presión adversaria, las llamadas "jailbreaks", probablemente exige capas de defensas más allá de las que normalmente se integran en el entrenamiento.

Crucialmente, esto romperá la correlación entre la capacidad y el motivo: el solitario perturbado que quiere matar a la gente pero carece de la disciplina o

habilidad para hacerlo ahora se elevará al nivel de capacidad del virólogo doctoral, que es poco probable que tenga esta motivación. Esta preocupación se generaliza más allá de la biología (aunque creo que la biología es el área más aterradora) a cualquier área donde la gran destrucción es posible, pero actualmente requiere un alto nivel de habilidad y disciplina. Para decirlo de otra manera, alquilar una poderosa IA da inteligencia a personas maliciosas (pero de otra manera promedio). Me preocupa que haya un gran número de tales personas por ahí, y que si tienen acceso a una manera fácil de matar a millones de personas, tarde o temprano uno de ellos lo hará. Además, aquellos que *do* tienen experiencia pueden estar capacitados para cometer una destrucción a mayor escala de lo que podían antes.

La biología es, con mucho, el área que más me preocupa, debido a su gran potencial de destrucción y la dificultad de defenderme contra ella, por lo que me centraré en la biología en particular. Pero gran parte de lo que digo aquí se aplica a otros riesgos, como los ataques cibernéticos, las armas químicas o la tecnología nuclear.

No voy a entrar en detalles sobre cómo fabricar armas biológicas, por razones que deberían ser obvias. Pero a un alto nivel, me preocupa que los LLM se estén acercando (o ya hayan alcanzado) el conocimiento necesario para crearlos y liberarlos de extremo a extremo, y que su potencial de destrucción sea muy alto. Algunos agentes biológicos podrían causar millones de muertes si se hiciera un esfuerzo decidido para liberarlos para una máxima propagación. Sin embargo, esto todavía tomaría un nivel muy alto de habilidad, incluyendo una serie de pasos y procedimientos muy específicos que no son ampliamente conocidos. Mi preocupación no es meramente un conocimiento fijo o estático. Me preocupa que los LLM puedan llevar a alguien de conocimiento y capacidad promedio y guiarlos a través de un proceso complejo que de otra manera podría salir mal o requerir la depuración de una manera interactiva, similar a cómo el soporte técnico podría ayudar a una persona no técnica a depurar y solucionar problemas complicados relacionados con la computadora (aunque este sería un proceso más extendido, probablemente durando semanas o meses).

Los LLM más capaces (sustancialmente más allá del poder de la actual) podrían ser capaces de permitir actos aún más aterradores. En 2024, un grupo de científicos prominentes escribió una carta advirtiendo sobre los riesgos de investigar y potencialmente crear un nuevo tipo de organismo peligroso: "vida de espejo". El ADN, el ARN, los ribosomas y las proteínas que componen los organismos biológicos tienen la misma quiralidad (también llamada "mano") que hace que no sean equivalentes a una versión de sí mismos reflejada en el espejo (al igual que su mano derecha no se puede girar de tal manera que sea idéntica a su izquierda). Pero todo el sistema de proteínas que se unen entre sí, la maquinaria de la síntesis de ADN y la traducción de ARN y la construcción y descomposición de las proteínas, todo depende de esta mano. Si los científicos hicieran versiones de este material biológico con la mano opuesta, y hay algunas ventajas potenciales de estos, como los medicamentos que duran más tiempo en el cuerpo, podría ser extremadamente peligroso. Esto se debe a que la vida de la mano izquierda, si se hiciera en forma de organismos completos capaces de reproducción (lo que sería muy difícil), sería potencialmente indigerible para cualquiera de los sistemas que descomponen el material biológico en la tierra, tendría una "clave" que no

encajaría en la "cerradura" de ninguna enzima existente. Esto significaría que podría proliferar de una manera incontrolable y desplazar toda la vida en el planeta, en el peor de los casos incluso destruyendo toda la vida en la tierra.

Existe una incertidumbre científica sustancial sobre la creación y los efectos potenciales de la vida espejo. La carta de 2024 acompañó un informe que concluyó que "las bacterias espejo podrían crearse de manera plausible en las próximas o pocas décadas", que es una amplia gama. Pero un modelo de IA lo suficientemente poderoso (para ser claros, mucho más capaces que cualquiera que tengamos hoy en día) podría ser capaz de descubrir cómo crearlo mucho más rápidamente, y realmente ayudar a alguien a hacerlo.

My view is that even though these are obscure risks, and might seem unlikely, the magnitude of the consequences is so large that they should be taken seriously as a first-class risk of AI systems.

Los escépticos han planteado una serie de objeciones a la gravedad de estos riesgos biológicos de los LLM, con los que no estoy de acuerdo, pero que vale la pena abordar. La mayoría cae en la categoría de no apreciar la trayectoria exponencial en la que está la tecnología. En 2023, cuando comenzamos a hablar sobre los riesgos biológicos de los LLM, los escépticos dijeron que toda la información necesaria estaba disponible en Google y los LLM no agregaron nada más allá de esto. Nunca fue cierto que Google podría darle toda la información necesaria: los genomas están disponibles libremente, pero como dije anteriormente, no se pueden obtener de esa manera ciertos pasos clave, así como una gran cantidad de conocimientos prácticos. Pero también, a finales de 2023, los LLM proporcionaban claramente información más allá de lo que Google podría dar para algunos pasos del proceso.

Después de esto, los escépticos se retiraron a la objeción de que los LLM no eran útiles *de extremo a extremo*, y no pudieron ayudar con la *adquisición* de armas biológicas en lugar de solo proporcionar información teórica. A partir de mediados de 2025, nuestras mediciones muestran que los LLM ya pueden estar proporcionando un aumento sustancial en varias áreas relevantes, tal vez duplicando o triplicando la probabilidad de éxito. Esto nos llevó a decidir que los modelos Claude Opus 4 (y los modelos posteriores de Sonnet 4.5, Opus 4.1 y Opus 4.5) debían ser lanzados bajo nuestras protecciones de nivel de seguridad de IA 3 en nuestro marco de política de escalamiento responsable, y para implementar salvaguardas contra este riesgo (más sobre esto más adelante). Creemos que los modelos probablemente ahora se están acercando al punto en el que, sin salvaguardas, podrían ser útiles para permitir que alguien con un título de STEM, pero no específicamente un título de biología, pase por todo el proceso de producción de un arma biológica.

Another objection is that there are other actions unrelated to AI that society can take to block the production of bioweapons. Most prominently, the gene synthesis industry makes biological specimens on demand, and there is no federal requirement that providers screen orders to make sure they do not contain pathogens. An MIT study found that 36 out of 38 providers fulfilled an order containing the sequence of the 1918 flu. I am supportive of mandated gene

synthesis screening that would make it harder for individuals to weaponize pathogens, in order to reduce both AI-driven biological risks and also biological risks in general. But this is not something we have today. It would also be only one tool in reducing risk; it is a complement to guardrails on AI systems, not a substitute.

La mejor objeción es una que rara vez he visto planteada: que hay una brecha entre los modelos que son útiles en principio y la propensión real de los malos actores a usarlos. La mayoría de los malos actores individuales son individuos perturbados, por lo que casi por definición su comportamiento es impredecible e irracional, y es *these* Los malos actores, los no calificados, que podrían haber podido beneficiarse más de la IA, lo que facilita mucho la muerte de muchas personas.[24]

[24] Un fenómeno extraño relacionado con los asesinos en masa es que el estilo de asesinato que eligen funciona casi como una moda grotesca. En los años setenta y ochenta, los asesinos en serie eran muy comunes, y los nuevos asesinos en serie a menudo copiaban el comportamiento de los asesinos en serie más establecidos o famosos. En los años noventa y 2000s, los tiroteos masivos se volvieron más comunes, mientras que los asesinos en serie se volvieron menos comunes. No hay un cambio tecnológico que desencadenó estos patrones de comportamiento, solo parece que los asesinos violentos estaban copiando el comportamiento de los demás y la cosa "popular" para copiar cambió.
El hecho de que un tipo de ataque violento sea posible, no significa que alguien decida hacerlo. Tal vez los ataques biológicos no serán atractivos porque es razonablemente probable que infecten al perpetrador, no atienden a las fantasías de estilo militar que muchos individuos o grupos violentos tienen, y es difícil atacar selectivamente a personas específicas. También podría ser que pasar por un proceso que lleva meses, incluso si una IA lo guía a través de él, implica una cantidad de paciencia que la mayoría de las personas perturbadas simplemente no tienen. Es posible que simplemente tengamos suerte y el motivo y la habilidad no se combinan, en la práctica, de la manera correcta.

But this seems like very flimsy protection to rely on. The motives of disturbed loners can change for any reason or no reason, and in fact there are already instances of LLMs being used in attacks (just not with biology). The focus on disturbed loners also ignores ideologically motivated terrorists, who are often willing to expend large amounts of time and effort (for example, the 9/11 hijackers). Wanting to kill as many people as possible is a motive that will probably arise sooner or later, and it unfortunately suggests bioweapons as the method. Even if this motive is extremely rare, it only has to materialize once. And as biology advances (increasingly driven by AI itself), it may also become possible to carry out more selective attacks (for example, targeted against people with specific ancestries), which adds yet another, very chilling, possible motive.

No creo que los ataques biológicos se lleven a cabo necesariamente en el instante en que sea ampliamente posible hacerlo; de hecho, apostaría en contra de eso. Pero sumados en millones de personas y algunos años de tiempo, creo que existe un grave riesgo de un ataque importante, y las consecuencias serían tan graves (con

víctimas potencialmente en millones o más) que creo que no tenemos más remedio que tomar medidas serias para prevenirlo.

Defensas

That brings us to how to defend against these risks. Here I see three things we can do. First, AI companies can put guardrails on their models to prevent them from helping to produce bioweapons. Anthropic is very actively doing this. Claude's Constitution, which mostly focuses on high-level principles and values, has a small number of specific hard-line prohibitions, and one of them relates to helping with the production of biological (or chemical, or nuclear, or radiological) weapons. But all models Puede ser jailbreak, and so as a second line of defense, we've implemented (since mid-2025, when our tests showed our models were starting to get close to the threshold where they might begin to pose a risk) a classifier that specifically detects and blocks bioweapon-related outputs. We regularly upgrade and improve these classifiers, and have generally found them highly robust even against sophisticated adversarial attacks.[25]

[25] Los despilfarradores casuales a veces creen que han comprometido estos clasificadores cuando consiguen que el modelo produzca una pieza específica de información, como la secuencia del genoma de un virus. Pero como expliqué antes, el modelo de amenaza que nos preocupa implica un consejo interactivo paso a paso que se extiende a lo largo de semanas o meses sobre pasos oscuros específicos en el proceso de producción de armas biológicas, y esto es contra lo que nuestros clasificadores pretenden defenderse. (A menudo describimos nuestra investigación como la búsqueda de jailbreaks "universales", que no solo funcionan en un contexto específico o estrecho, sino que abren ampliamente el comportamiento del modelo).
These classifiers increase the costs to serve our models measurably (in some models, they are close to 5% of total inference costs) and thus cut into our margins, but we feel that using them is the right thing to do.

Para su crédito, algunas otras compañías de IA have implemented classifiers as well. But not every company has, and there is also nothing requiring companies to keep their classifiers. I am concerned that over time there may be a prisoner's dilemma where companies can defect and lower their costs by removing classifiers. This is once again a classic negative externalities problem that can't be solved by the voluntary actions of Anthropic or any other single company alone.[26]

[26] Aunque seguiremos invirtiendo en el trabajo para hacer que nuestros clasificadores sean más eficientes, y puede tener sentido que las empresas compartan avances como estos entre sí.
Voluntary industry standards may help, as may third-party evaluations and verification of the type done by AI security institutes y third-party evaluators.

But ultimately defense may require government action, which is the second thing we can do. My views here are the same as they are for addressing autonomy risks: we should start with transparency requirements,[27]

[27] **Obviously, I do not think companies should have to disclose technical details about the specific steps in biological weapons production that they are blocking, and the transparency legislation that has been passed so far (SB 53 and RAISE) accounts for this issue.**
which help society measure, monitor, and collectively defend against risks without disrupting economic activity in a heavy-handed way. Then, if and when we reach clearer thresholds of risk, we can craft legislation that more precisely targets these risks and has a lower chance of collateral damage. In the particular case of bioweapons, I actually think that the time for such targeted legislation may be approaching soon—Anthropic and other companies are learning more and more about the nature of biological risks and what is reasonable to require of companies in defending against them. Fully defending against these risks may require working internationally, even with geopolitical adversaries, but there is precedent in treaties prohibiting the development of biological weapons. I am generally a skeptic about most kinds of international cooperation on AI, but this may be one narrow area where there is some chance of achieving global restraint. Even dictatorships do not want massive bioterrorist attacks.

Finalmente, la tercera contramedida que podemos tomar es tratar de desarrollar defensas contra los ataques biológicos. Esto podría incluir el monitoreo y seguimiento de la detección temprana, inversiones en I+D de purificación de aire (por ejemplo far-UVC disinfection), rapid vaccine development that can respond and adapt to an attack, better personal protective equipment (PPE),[28]

[28] **Another related idea is "resilience markets" where the government encourages stockpiling of PPE, respirators, and other essential equipment needed to respond to a biological attack by promising ahead of time to pay a pre-agreed price for this equipment in an emergency. This incentivizes suppliers to stockpile such equipment without fear that the government will seize it without compensation.**
and treatments or vaccinations for some of the most likely biological agents. mRNA vaccines, which can be designed to respond to a particular virus or variant, are an early example of what is possible here. Anthropic is excited to work with biotech and pharmaceutical companies on this problem. But unfortunately I think our expectations on the defensive side should be limited. There is an asymmetry between attack and defense in biology, because agents spread rapidly on their own, while defenses require detection, vaccination, and treatment to be organized across large numbers of people very quickly in response. Unless the response is lightning quick (which it rarely is), much of the damage will be done before a response is possible. It is conceivable that future technological improvements could shift this balance in favor of defense (and we should certainly use AI to help develop such technological advances), pero hasta entonces, las salvaguardias preventivas serán nuestra principal línea de defensa.

It's worth a brief mention of cyberattacks here, since unlike biological attacks, AI-led cyberattacks have actually happened in the wild, including at a large scale and for state-sponsored espionage. We expect these attacks to become more capable as models advance rapidly, until they are the main way in which cyberattacks are conducted. I expect AI-led cyberattacks to become a serious and unprecedented threat to the integrity of computer systems around the world, and Anthropic is working very hard to shut down these attacks and eventually reliably prevent

them from happening. The reason I haven't focused on cyber as much as biology is that (1) cyberattacks are much less likely to kill people, certainly not at the scale of biological attacks, and (2) the offense-defense balance may be more tractable in cyber, where there is at least some hope that defense could keep up with (and even ideally outpace) AI attack if we invest in it properly.

Aunque la biología es actualmente el vector de ataque más grave, hay muchos otros vectores y es posible que pueda surgir uno más peligroso. El principio general es que sin contramedidas, es probable que la IA reduzca continuamente la barrera de la actividad destructiva a una escala cada vez mayor, y la humanidad necesita una respuesta seria a esta amenaza.

3. El aparato odioso

Misuse for seizing power

The previous section discussed the risk of individuals and small organizations co-opting a small subset of the "country of geniuses in a datacenter" to cause large-scale destruction. But we should also worry—likely substantially more so—about misuse of AI for the purpose of *wielding or seizing power*, likely by larger and more established actors.[29]

[29] ¿Por qué estoy más preocupado por los grandes actores por tomar el poder, pero los pequeños actores por causar destrucción? Porque la dinámica es diferente. Aprovechar el poder se trata de si un actor puede acumular suficiente fuerza para superar a todos los demás, por lo que deberíamos preocuparnos por los actores más poderosos y/o los más cercanos a la IA. La destrucción, por el contrario, puede ser causada por aquellos con poco poder si es mucho más difícil defenderse que causar. Es entonces un juego de defensa contra las amenazas más *numerosas*, que probablemente serán actores más pequeños.

En *Machines of Loving Grace*, discutí la posibilidad *de* que los gobiernos autoritarios pudieran usar una poderosa IA para vigilar o reprimir a sus ciudadanos de maneras que serían extremadamente difíciles de reformar o derrocar. Las autocracias actuales están limitadas en lo represivas que pueden ser por la necesidad de que los humanos lleven a cabo sus órdenes, y los humanos a menudo tienen límites en lo inhumanos que están dispuestos a ser. Pero las autocracias habilitadas para la IA no tendrían tales límites.

Peor aún, los países también podrían usar su ventaja en la inteligencia artificial para ganar poder sobre *otros países*. Si el "país de los genios" en su conjunto fuera simplemente propiedad y controlado por el aparato militar de un solo país (humano), y otros países no tuvieran capacidades equivalentes, es difícil ver cómo podrían defenderse: serían burlados a cada paso, similar a una guerra entre humanos y ratones. La unión de estas dos preocupaciones conduce a la alarmante posibilidad de una dictadura totalitaria global. Obviamente, debería ser una de nuestras principales prioridades para evitar este resultado.

Hay muchas maneras en que la IA podría habilitar, afianzar o expandir la autocracia, pero enumeraré algunas de las que más me preocupa. Tenga en cuenta

que algunas de estas aplicaciones tienen usos legítimos defensivos, y no estoy necesariamente argumentando en contra de ellas en términos absolutos; sin embargo, me preocupa que estructuralmente tiendan a favorecer las autocracias:

- **Armas totalmente autónomas. Un enjambre de millones o miles de millones de drones armados totalmente automatizados, controlados localmente por una poderosa inteligencia artificial y coordinados estratégicamente en todo el mundo por una IA aún más poderosa, podría ser un ejército inmejorable, capaz de derrotar a cualquier ejército en el mundo y suprimir la disidencia dentro de un país siguiendo a todos los ciudadanos. Los desarrollos en la Guerra Rusia-Ucrania deberían alertarnos sobre el hecho de que la guerra de drones ya está con nosotros (aunque aún no es completamente autónoma, y una pequeña fracción de lo que podría ser posible con una poderosa IA). La I + D de la poderosa IA podría hacer que los drones de un país sean muy superiores a los de otros, acelerar su fabricación, hacerlos más resistentes a los ataques electrónicos, mejorar sus maniobras, etc. Por supuesto, estas armas también tienen usos legítimos en la defensa de la democracia: han sido clave para defender a Ucrania y probablemente serían clave para defender a Taiwán. Pero son un arma peligrosa para manejar: debemos preocuparnos por ellos en manos de las autocracias, pero también preocuparnos de que, debido a que son tan poderosos, con tan poca responsabilidad, existe un riesgo mucho mayor de que los gobiernos democráticos los vuelvan en contra de su propio pueblo para tomar el poder.**
- **Vigilancia de la IA. Una IA suficientemente poderosa probablemente podría usarse para comprometer cualquier sistema informático en el mundo,[30]**

  30 Esto puede parecer que está en tensión con mi punto de que el ataque y la defensa pueden estar más equilibrados con los ciberataques que con las armas biológicas, pero mi preocupación aquí es que si la IA de un país es la más poderosa del mundo, entonces otros no podrán defender incluso si la tecnología en sí tiene un equilibrio de ataque-defensa intrínseca.

  Y también podría utilizar el acceso obtenido de esta manera para leer *Y dar sentido a* todas las comunicaciones electrónicas del mundo (o incluso todas las comunicaciones en persona del mundo, si los dispositivos de grabación se pueden construir o requisar). Podría ser aterradoramente plausible simplemente generar una lista completa de cualquier persona que no esté de acuerdo con el gobierno en cualquier número de temas, incluso si tal desacuerdo no es explícito en nada que digan o hagan. Una poderosa inteligencia artificial que analiza miles de millones de conversaciones de millones de personas podría medir el sentimiento público, detectar los focos de deslealtad que se forman y eliminarlos antes de que crezcan. Esto podría llevar a la imposición de un verdadero panóptico en una escala que no vemos hoy, incluso con el PCCh.

- **La propaganda de la IA. Los fenómenos actuales de "psicosis de IA" y "novias de IA" sugieren que incluso en su nivel actual de inteligencia, los modelos de IA pueden tener una poderosa influencia psicológica en las personas. Las versiones mucho más poderosas de estos modelos, que**

estaban mucho más incrustados y conscientes de la vida diaria de las personas y podían modelarlas e influir en ellos durante meses o años, probablemente serían capaces de esencialmente lavar el cerebro a muchos (¿la mayoría?) La gente en cualquier ideología o actitud deseada, y podría ser empleada por un líder sin escrúpulos para garantizar la lealtad y suprimir la disidencia, incluso frente a un nivel de represión contra el que la mayoría de las poblaciones se rebelarían. Hoy en día la gente se preocupa mucho, por ejemplo, por la influencia potencial de TikTok como propaganda del PCCh dirigida a los niños. También me preocupa eso, pero un agente de inteligencia artificial personalizado que lo conozca a lo largo de los años y use su conocimiento de usted para dar forma a todas sus opiniones sería dramáticamente más poderoso que esto.

- **Toma de decisiones estratégicas. Un país de genios en un centro de datos podría usarse para asesorar a un país, grupo o individuo sobre estrategia geopolítica, lo que podríamos llamar un "Bismarck virtual". Podría optimizar las tres estrategias anteriores para tomar el poder, además de desarrollar probablemente muchas otras en las que no he pensado (pero que un país de genios podría). Es probable que la diplomacia, la estrategia militar, la I + D, la estrategia económica y muchas otras áreas aumenten sustancialmente su eficacia mediante una poderosa IA. Muchas de estas habilidades serían legítimamente útiles para las democracias, queremos que las democracias tengan acceso a las mejores estrategias para defenderse contra las autocracias, pero el potencial de mal uso en las manos de *cualquiera* sigue siendo.**

Habiendo descrito *lo* que me preocupa, pasemos a *quién*. Me preocupan las entidades que tienen más acceso a la IA, que están partiendo de una posición del poder más político, o que tienen una historia de represión existente. En orden de severidad, me preocupa:

- **El PCCh. China es el segundo después de los Estados Unidos en capacidades de inteligencia artificial, y es el país con la mayor probabilidad de superar a los Estados Unidos en esas capacidades. Su gobierno es actualmente autocrático y opera un estado de vigilancia de alta tecnología. Ya ha desplegado vigilancia basada en la inteligencia artificial (incluso en la represión de los uigures), y se cree que emplea propaganda algorítmica a través de TikTok (además de sus muchos otros esfuerzos de propaganda internacional). Tienen las manos en el camino más claro hacia la pesadilla totalitaria habilitada para la IA que presenté arriba. Incluso puede ser el resultado predeterminado dentro de China, así como dentro de otros estados autocráticos a quienes el PCCh exporta tecnología de vigilancia. He escrito a menudo sobre la amenaza de que el PCCh tome la iniciativa en la IA y el imperativo existencial para evitar que lo hagan. Esta es la razón. Para ser claros, no estoy señalando a China por animus en particular, son simplemente el país que más combina la destreza de la IA, un gobierno autocrático y un estado de vigilancia de alta tecnología. En todo caso, es el propio pueblo chino el que tiene más probabilidades de sufrir la represión habilitada por la IA del PCCh, y no tienen voz en las acciones de su gobierno. Admiro y respeto mucho al pueblo chino y apoyo a los muchos disidentes valientes dentro de China y su lucha por la libertad.**

- **Democracies competitive in AI. As I wrote above, democracies have a legitimate interest in some AI-powered military and geopolitical tools, because democratic governments offer the best chance to counter the use of these tools by autocracies. Broadly, I am supportive of arming democracies with the tools needed to defeat autocracies in the age of AI—I simply don't think there is any other way. But we cannot ignore the potential for abuse of these technologies by democratic governments themselves. Democracies normally have safeguards that prevent their military and intelligence apparatus from being turned inwards against their own population,[31]**

  31 For example, in the United States this includes the fourth amendment and the Posse Comitatus Act.

  but because AI tools require so few people to operate, there is potential for them to circumvent these safeguards and the norms that support them. It is also worth noting that some of these safeguards are already gradually eroding in some democracies. Thus, we should arm democracies with AI, but we should do so carefully and within limits: they are the immune system we need to fight autocracies, but like the immune system, there is some risk of them turning on us and becoming a threat themselves.

- **Non-democratic countries with large datacenters. Beyond China, most countries with less democratic governance are not leading AI players in the sense that they don't have companies which produce frontier AI models. Thus they pose a fundamentally different and lesser risk than the CCP, which remains the primary concern (most are also less repressive, and the ones that are more repressive, like North Korea, have no significant AI industry at all). But some of these countries do have large *datacenters* (often as part of buildouts by companies operating in democracies), which can be used to run frontier AI at large scale (though this does not confer the ability to push the frontier). There is some amount of danger associated with this—these governments could in principle expropriate the datacenters and use the country of AIs within it for their own ends. I am less worried about this compared to countries like China that directly develop AI, but it's a risk to keep in mind.[32]**

  32 Also, to be clear, there are some arguments for building large datacenters in countries with varying governance structures, particularly if they are controlled by companies in democracies. Such buildouts could in principle help democracies compete better with the CCP, which is the greater threat. I also think such datacenters don't pose much risk unless they are very large. But on balance, I think caution is warranted when placing very large datacenters in countries where institutional safeguards and rule-of-law protections are less well-established.

- **AI companies. It is somewhat awkward to say this as the CEO of an AI company, but I think the next tier of risk is actually AI companies themselves. AI companies control large datacenters, train frontier models, have the greatest expertise on how to use those models, and in some cases have daily contact with and the possibility of influence over tens or hundreds of millions of users. The main thing they lack is the legitimacy and infrastructure of a state, so much of what would be needed to build the tools of an AI autocracy would be illegal for an AI company to do, or at least exceedingly suspicious. But some of it is not impossible: they could, for**

example, use their AI products to brainwash their massive consumer user base, and the public should be alert to the risk this represents. I think the governance of AI companies deserves a lot of scrutiny.

There are a number of possible arguments against the severity of these threats, and I wish I believed them, because AI-enabled authoritarianism terrifies me. It's worth going through some of these arguments and responding to them.

En primer lugar, algunas personas podrían poner su fe en la disuasión nuclear, particularmente para contrarrestar el uso de armas autónomas de IA para la conquista militar. Si alguien amenaza con usar estas armas contra usted, siempre puede amenazar una respuesta nuclear. Mi preocupación es que estoy not totally sure we can be confident in the nuclear deterrent against a country of geniuses in a datacenter: it is possible that powerful AI could devise ways to detect and strike nuclear submarines, conduct influence operations against the operators of nuclear weapons infrastructure, or use AI's cyber capabilities to launch a cyberattack against satellites used to detect nuclear launches.[33]

[33] Este es, por supuesto, también un argumento para mejorar la seguridad de la disuasión nuclear para que sea más probable que sea robusta contra la poderosa IA, y las democracias con armas nucleares deberían hacerlo. Pero no sabemos de qué será capaz una poderosa IA o qué defensas, si las hay, funcionarán en contra, por lo que no debemos asumir que estas medidas necesariamente resolverán el problema.

Alternatively, it's possible that taking over countries is feasible with only AI surveillance and AI propaganda, and never actually presents a clear moment where it's obvious what is going on and where a nuclear response would be appropriate. *Maybe* these things aren't feasible and the nuclear deterrent will still be effective, but it seems too high stakes to take a risk.[34]

34 There is also the risk that even if the nuclear deterrent remains effective, an attacking country might decide to call our bluff—it's unclear whether we'd be willing to use nuclear weapons to defend against a drone swarm even if the drone swarm has a substantial risk of conquering us. Drone swarms might be a new thing that is less severe than nuclear attacks but more severe than conventional attacks. Alternatively, differing assessments of the effectiveness of the nuclear deterrent in the age of AI might alter the game theory of nuclear conflict in a destabilizing manner.

Una segunda posible objeción es que puede haber contramedidas que podamos tomar contra estas herramientas de la autocracia. Podemos contrarrestar los drones con nuestros propios drones, la ciberdefensa mejorará junto con el ciberataque, puede haber formas de inmunizar a las personas contra la propaganda, etc. Mi respuesta es que estas defensas solo serán posibles con una IA comparablemente poderosa. Si no hay alguna contrafuerza con un país comparablemente inteligente y numerosos genios en un centro de datos, no será posible igualar la calidad o la cantidad de drones, para la ciberdefensa para superar la ciberofensa, etc. Así que la cuestión de las contramedidas se reduce a la cuestión de un equilibrio de poder en la poderosa IA. Aquí, me preocupa la propiedad recursiva o auto-reforzada de la poderosa IA (que discutí al principio de este ensayo): que cada generación de IA se puede utilizar para diseñar y entrenar a la próxima generación de IA. Esto conduce a un riesgo de una ventaja descontrolada, donde el líder actual en inteligencia artificial poderosa puede ser

capaz de aumentar su liderazgo y puede ser difícil de alcanzar. Tenemos que asegurarnos de que no sea un país autoritario el que llegue a este ciclo primero.

Además, incluso si se puede lograr un equilibrio de poder, todavía existe el riesgo de que el mundo se divida en esferas autocráticas, como en *mil novecientos ochenta y cuatro*. Incluso si varios poderes competitivos tienen sus poderosos modelos de IA, y ninguno puede dominar a los demás, cada poder todavía podría reprimir internamente a su propia población, y sería muy difícil de derrocar (ya que las poblaciones no tienen una IA poderosa para defenderse). Por lo tanto, es importante prevenir la autocracia habilitada para la IA, incluso si no conduce a que un solo país se apodere del mundo.

Defensas

How do we defend against this wide range of autocratic tools and potential threat actors? As in the previous sections, there are several things I think we can do. First, we should absolutely not be selling chips, chip-making tools, or datacenters to the CCP. Chips and chip-making tools are the single greatest bottleneck to powerful AI, and blocking them is a simple but extremely effective measure, perhaps the most important single action we can take. It makes no sense to sell the CCP the tools with which to build an AI totalitarian state and possibly conquer us militarily. A number of complicated arguments are made to justify such sales, such as the idea that "spreading our tech stack around the world" allows "America to win" in some general, unspecified economic battle. In my view, this is like selling nuclear weapons to North Korea and then bragging that the missile casings are made by Boeing and so the US is "winning." China is several years behind the US in their ability to produce frontier chips in quantity, and the critical period for building the country of geniuses in a datacenter is very likely to be within those next several years.[35]

[35] Para ser claros, creo que es la estrategia correcta no vender chips a China, incluso si la línea de tiempo para una IA poderosa fuera sustancialmente más larga. No podemos hacer que los chinos sean "adictos" a los chips estadounidenses, están decididos a desarrollar su industria de chips nativos de una manera u otra. Les llevará muchos años hacerlo, y todo lo que estamos haciendo vendiendo esos chips les está dando un gran impulso durante ese tiempo.
There is no reason to give a giant boost to their AI industry during this critical period.

En segundo lugar, tiene sentido utilizar la inteligencia artificial para empoderar a las democracias para resistir las autocracias. Esta es la razón por la que Anthropic considera importante proporcionar IA a las comunidades de inteligencia y defensa en los Estados Unidos y sus aliados democráticos. La defensa de las democracias que están bajo ataque, como Ucrania y (a través de ataques cibernéticos) Taiwán, parece una prioridad especialmente alta, al igual que empoderar a las democracias para que utilicen sus servicios de inteligencia para interrumpir y degradar las autocracias desde el interior. En algún nivel, la única manera de responder a las amenazas autocráticas es igualarlas y superarlas militarmente. Una coalición de Estados Unidos y sus aliados democráticos, si lograra el predominio en una

poderosa IA, estaría en condiciones no solo de defenderse de las autocracias, sino de contenerlas y limitar sus abusos totalitarios de la IA.

Third, we need to draw a hard line against AI abuses within democracies. There need to be limits to what we allow our governments to do with AI, so that they don't seize power or repress their own people. The formulation I have come up with is that we should use AI for national defense in all ways *except those which would make us more like our autocratic adversaries*.

¿Dónde se debe trazar la línea? En la lista al comienzo de esta sección, dos elementos, utilizando la inteligencia artificial para la vigilancia masiva doméstica y la propaganda masiva, me parecen líneas rojas brillantes y completamente ilegítimas. Algunos podrían argumentar que no hay necesidad de hacer nada (al menos en los Estados Unidos), ya que la vigilancia masiva doméstica ya es ilegal bajo la Cuarta Enmienda. Pero el rápido progreso de la IA puede crear situaciones que nuestros marcos legales existentes no están bien diseñados para tratar. Por ejemplo, es probable que no sea inconstitucional que el gobierno de los Estados Unidos realice grabaciones a escala masiva de todas las *conversaciones* públicas (por ejemplo, cosas que la gente se dice entre sí en una esquina de la calle), y anteriormente habría sido difícil clasificar este volumen de información, pero con la IA podría transcribirse, interpretarse y triangularse para crear una imagen de la actitud y lealtades de muchos o la mayoría de los ciudadanos. Apoyaría la legislación centrada en las libertades civiles (o tal vez incluso una enmienda constitucional) que imponga barandillas más fuertes contra los abusos impulsados por la IA.

The other two items—fully autonomous weapons and AI for strategic decision-making—are harder lines to draw since they have legitimate uses in defending democracy, while also being prone to abuse. Here I think what is warranted is extreme care and scrutiny combined with guardrails to prevent abuses. My main fear is having too small a number of "fingers on the button," such that one or a handful of people could essentially operate a drone army without needing any other humans to cooperate to carry out their orders. As AI systems get more powerful, we may need to have more direct and immediate oversight mechanisms to ensure they are not misused, perhaps involving branches of government other than the executive. I think we should approach fully autonomous weapons in particular with great caution,[36]

[36] To be clear, most of what is being used in Ukraine and Taiwan today are not *fully* autonomous weapons. These are coming, but not here today.
and not rush into their use without proper safeguards.

En cuarto lugar, después de trazar una línea dura contra los abusos de la IA en las democracias, deberíamos usar ese precedente para crear un tabú internacional contra los peores abusos de la poderosa IA. Reconozco que los vientos políticos actuales se han vuelto contra la cooperación internacional y las normas internacionales, pero este es un caso en el que los necesitamos profundamente. El mundo necesita entender el oscuro potencial de la poderosa IA en manos de los autócratas, y reconocer que ciertos usos de la IA equivalen a un intento de robar permanentemente su libertad e imponer un estado totalitario del que no pueden

escapar. Incluso diría que en algunos casos, la vigilancia a gran escala con una poderosa IA, la propaganda masiva con una poderosa IA y ciertos tipos de usos *ofensivos* de armas totalmente autónomas deberían considerarse crímenes contra la humanidad. En términos más generales, se necesita una norma robusta contra el totalitarismo habilitado por la IA y todas sus herramientas e instrumentos.

It is possible to have an even stronger version of this position, which is that because the possibilities of AI-enabled totalitarianism are so dark, autocracy is simply not a form of government that people can accept in the post-powerful AI age. Just as feudalism became unworkable with the industrial revolution, the AI age could lead inevitably and logically to the conclusion that democracy (and, hopefully, democracy improved and reinvigorated by AI, as I discuss in *Machines of Loving Grace*) is the only viable form of government if humanity is to have a good future.

Fifth and finally, AI companies should be carefully watched, as should their connection to the government, which is necessary, but must have limits and boundaries. The sheer amount of capability embodied in powerful AI is such that ordinary corporate governance—which is designed to protect shareholders and prevent ordinary abuses such as fraud—is unlikely to be up to the task of governing AI companies. There may also be value in companies publicly committing to (perhaps even as part of corporate governance) not take certain actions, such as privately building or stockpiling military hardware, using large amounts of computing resources by single individuals in unaccountable ways, or using their AI products as propaganda to manipulate public opinion in their favor.

El peligro aquí viene de muchas direcciones, y algunas direcciones están en tensión con otras. La única constante es que debemos buscar la rendición de cuentas, las normas y las barandillas para todos, incluso cuando empoderamos a los "buenos" actores para que mantengan bajo control a los actores "malos".

4. Piano del jugador

Economic disruption

Las tres secciones anteriores eran esencialmente sobre los riesgos de seguridad planteados por la poderosa IA: los riesgos de la propia IA, los riesgos del uso indebido por parte de individuos y pequeñas organizaciones y los riesgos de uso indebido por parte de los estados y las grandes organizaciones. Si dejamos de lado los riesgos de seguridad o asumimos que se han resuelto, la siguiente pregunta es económica. ¿Cuál será el efecto de esta infusión de increíble capital "humano" en la economía? Claramente, el efecto más obvio será aumentar en gran medida el crecimiento económico. El ritmo de los avances en la investigación científica, la innovación biomédica, la fabricación, las cadenas de suministro, la eficiencia del sistema financiero y mucho más casi están garantizados para conducir a una tasa de crecimiento económico mucho más rápida. En *Machines of Loving Grace*, sugiero que una tasa de crecimiento anual del PIB sostenida del 10-20% puede ser posible.

But it should be clear that this is a double-edged sword: what are the economic prospects for most existing humans in such a world? New technologies often bring

labor market shocks, and in the past humans have always recovered from them, but I am concerned that this is because these previous shocks affected only a small fraction of the full possible range of human abilities, leaving room for humans to expand to new tasks. AI will have effects that are much broader and occur much faster, and therefore I worry it will be much more challenging to make things work out well.

Disrupción del mercado laboral

Hay dos problemas específicos que me preocupan: el desplazamiento del mercado laboral y la concentración del poder económico. Empecemos por el primero. Este es un tema sobre el que advertí muy públicamente en 2025, donde predije que la IA podría desplazar a la mitad de todos los empleos de cuello blanco de nivel de entrada en los próximos 1 a 5 años, incluso cuando acelera el crecimiento económico y el progreso científico. Esta advertencia inició un debate público sobre el tema. Muchos CEOs, tecnólogos y economistas estuvieron de acuerdo conmigo, pero otros asumieron que estaba cayendo presa de una falacia "grumal de trabajo" y no sabían cómo funcionaban los mercados laborales, y algunos no vieron el rango de tiempo de 1 a 5 años y pensaron que estaba afirmando que la IA está desplazando empleos en este momento (lo que estoy de acuerdo en que probablemente no lo sea). Por lo tanto, vale la pena pasar en detalle por qué estoy preocupado por el desplazamiento laboral, para aclarar estos malentendidos.

Como línea de base, es útil entender cómo los mercados laborales *normalmente* responden a los avances en tecnología. Cuando aparece una nueva tecnología, comienza haciendo piezas de un trabajo humano determinado más eficiente. Por ejemplo, al principio de la Revolución Industrial, las máquinas, como los arados mejorados, permitieron a los agricultores humanos ser más eficientes en algunos aspectos del trabajo. Esto mejoró la productividad de los agricultores, lo que aumentó sus salarios.

En la siguiente etapa, algunas partes del trabajo de la agricultura podrían realizarse *completamente* por máquinas, por ejemplo, con la invención de la máquina de trilla o taladro de semillas. En esta fase, los humanos hicieron una fracción cada vez menor del trabajo, pero el trabajo que *realizaron* se volvió cada vez más apalancado porque es complementario al trabajo de las máquinas, y su productividad continuó aumentando. Como lo describe la paradoja de Jevons, los salarios de los agricultores y tal vez incluso el número de agricultores siguieron aumentando. Incluso cuando el 90% del trabajo está siendo realizado por máquinas, los humanos simplemente pueden hacer 10 veces más del 10% que todavía hacen, produciendo 10 veces más producción para la misma cantidad de mano de obra.

Finalmente, las máquinas hacen todo o casi todo, como con las cosechadoras, tractores y otros equipos modernos. En este punto, la agricultura como una forma de empleo humano realmente entra en fuerte declive, y esto potencialmente causa graves interrupciones en el corto plazo, pero debido a que la agricultura es solo una de las muchas actividades útiles que los humanos pueden hacer, las personas eventualmente cambian a otros trabajos, como operar máquinas de fábrica. Esto es cierto a pesar de que la agricultura representó una gran proporción del empleo

ex ante. Hace 250 años, el 90% de los estadounidenses vivían en granjas; en Europa, el 50-60% del empleo era agrícola. Ahora esos porcentajes están en los dígitos bajos en esos lugares, porque los trabajadores cambiaron a trabajos industriales (y más tarde, trabajos de trabajo de conocimiento). La economía puede hacer lo que antes requería la mayor parte de la fuerza de trabajo con solo el 1-2% de ella, liberando al resto de la fuerza de trabajo para construir una sociedad industrial cada vez más avanzada. No hay un "bulto de trabajo" fijo, solo una capacidad cada vez mayor para hacer más y más con cada vez menos. Los salarios de las personas aumentan en línea con el PIB exponencial y la economía mantiene el pleno empleo una vez que han pasado las interrupciones en el corto plazo.

It's possible things will go roughly the same way with AI, but I would bet pretty strongly against it. Here are some reasons I think AI is likely to be different:

- **Speed. The pace of progress in AI is much faster than for previous technological revolutions. For example, in the last 2 years, AI models went from barely being able to complete a single line of code, to writing all or almost all of the code for some people—including engineers at Anthropic.[37]**

  37 Our model card for Claude Opus 4.5, our most recent model, shows that Opus performs better on a performance engineering interview frequently given at Anthropic than any interviewee in the history of the company.

  Soon, they may do the entire task of a software engineer end to end.[38]

  38 "Writing all of the code" and "doing the task of a software engineer end to end" are very different things, because software engineers do much more than just write code, including testing, dealing with environments, files, and installation, managing cloud compute deployments, iterating on products, and much more.

  It is hard for people to adapt to this pace of change, both to the changes in how a given job works and in the need to switch to new jobs. Even legendary programmers are increasingly Describiéndose a sí mismos como "detrás". The pace may if anything continue to speed up, as AI coding models increasingly accelerate the task of AI development. To be clear, speed in itself does not mean labor markets and employment won't eventually recover, it just implies the short-term transition will be unusually painful compared to past technologies, since humans and labor markets are slow to react and to equilibrate.

- **Cognitive breadth. As suggested by the phrase "country of geniuses in a datacenter," AI will be capable of a very wide range of human cognitive abilities—perhaps all of them. This is very different from previous technologies like mechanized farming, transportation, or even computers.[39]**

  39 Las computadoras son generales en cierto sentido, pero son claramente incapaces por sí solas de la gran mayoría de las habilidades cognitivas humanas, incluso cuando exceden en gran medida a los humanos en unas pocas áreas (como la aritmética). Por supuesto, las cosas construidas *sobre* las computadoras, como la inteligencia artificial, ahora son capaces de una amplia gama de habilidades cognitivas, que es de lo que se trata este ensayo.

This will make it harder for people to switch easily from jobs that are displaced to similar jobs that they would be a good fit for. For example, the general intellectual abilities required for entry-level jobs in, say, finance, consulting, and law are fairly similar, even if the specific knowledge is quite different. A technology that disrupted only one of these three would allow employees to switch to the two other close substitutes (or for undergraduates to switch majors). But disrupting all three at once (along with many other similar jobs) may be harder for people to adapt to. Furthermore, it's not *Sólo* Que la mayoría de los puestos de trabajo existentes se verán afectados. Esa parte ha sucedido antes: recuerde que la agricultura era un gran porcentaje del empleo. Pero los agricultores podían cambiar al trabajo relativamente similar de operar máquinas de fábrica, a pesar de que ese trabajo no había sido común antes. Por el contrario, la IA coincide cada vez más con el perfil cognitivo general de los humanos, lo que significa que también será bueno en los nuevos trabajos que normalmente se crearían en respuesta a los antiguos que se están automatizando. Otra forma de decir que la IA no es un sustituto de trabajos humanos específicos, sino más bien un sustituto general del trabajo para los humanos.

- **Slicing by cognitive ability. A través de una amplia gama de tareas, la IA parece estar avanzando desde la parte inferior de la escalera de habilidades hasta la parte superior. Por ejemplo, en la codificación nuestros modelos han pasado del nivel de "un codificador mediocre" a "un codificador fuerte" a "un codificador muy fuerte".[40]**

40 To be clear, AI models do not have precisely the same profile of strengths and weaknesses as humans. But they are also advancing fairly uniformly along every dimension, such that having a spiky or uneven profile may not ultimately matter.

We are now starting to see the same progression in white-collar work in general. We are thus at risk of a situation where, instead of affecting people with specific skills or in specific professions (who can adapt by retraining), AI is affecting people with certain intrinsic cognitive properties, namely lower intellectual ability (which is harder to change). It is not clear where these people will go or what they will do, and I am concerned that they could form an unemployed or very-low-wage "underclass." To be clear, things somewhat like this have happened before—for example, computers and the internet are believed by some economists to represent "skill-biased technological change." But this skill biasing was both not as extreme as what I expect to see with AI, and is believed to have contributed to an increase in wage inequality,[41]

41 Aunque hay debate entre los economistas sobre esta idea.

Por lo tanto, no es exactamente un precedente tranquilizador.

- **Ability to fill in the gaps. The way human jobs often adjust in the face of new technology is that there are many aspects to the job, and the new technology, even if it appears to directly replace humans, often has gaps in it. If someone invents a machine to make widgets, humans may still have to**

> **load raw material into the machine. Even if that takes only 1% as much effort as making the widgets manually, human workers can simply make 100x more widgets. But AI, in addition to being a rapidly advancing technology, is also a rapidly *adapting* technology. During every model release, AI companies carefully measure what the model is good at and what it isn't, and customers also provide such information after the launch. Weaknesses can be addressed by collecting tasks that embody the current gap, and training on them for the next model. Early in generative AI, users noticed that AI systems had certain weaknesses (such as AI image models generating hands with the wrong number of fingers) and many assumed these weaknesses were inherent to the technology. If they were, it would limit job disruption. But pretty much every such weakness gets addressed quickly— often, within just a few months.**

Vale la pena abordar puntos comunes de escepticismo. En primer lugar, existe el argumento de que la difusión económica será lenta, de tal manera que incluso si la tecnología subyacente es *capaz* de hacer la mayor parte del trabajo humano, la aplicación real de la misma en toda la economía puede ser mucho más lenta (por ejemplo, en industrias que están lejos de la industria de la IA y lentas de adoptar). La difusión lenta de la tecnología es definitivamente real: hablo con personas de una amplia variedad de empresas, y hay lugares donde la adopción de la IA tomará años. Es por eso que mi predicción para el 50% de los trabajos de cuello blanco de nivel de entrada que se interrumpen es de 1 a 5 años, a pesar de que sospecho que tendremos una IA poderosa (que sería, tecnológicamente hablando, suficiente para hacer *la mayoría o todos* los trabajos, no solo el nivel de entrada) en mucho menos de 5 años. Pero los efectos de difusión simplemente nos dan tiempo. Y no estoy seguro de que serán tan lentos como la gente predice. La adopción de la IA empresarial está creciendo a un ritmo mucho más rápido que cualquier tecnología anterior, en gran medida con la fuerza pura de la tecnología en sí. Además, incluso si las empresas tradicionales tardan en adoptar nuevas tecnologías, las nuevas empresas surgirán para servir como "pegamento" y facilitar la adopción. Si eso no funciona, las startups pueden simplemente interrumpir directamente a los titulares.

That could lead to a world where it isn't so much that specific jobs are disrupted as it is that large enterprises are disrupted in general and replaced with much less labor-intensive startups. This could also lead to a world of "geographic inequality," where an increasing fraction of the world's wealth is concentrated in Silicon Valley, which becomes its own economy running at a different speed than the rest of the world and leaving it behind. All of these outcomes would be great for economic growth—but not so great for the labor market or those who are left behind.

En segundo lugar, algunas personas dicen que los empleos humanos se trasladarán al mundo físico, lo que evita toda la categoría de "trabajo cognitivo" donde la IA está progresando tan rápidamente. Tampoco estoy seguro de qué tan seguro es esto. Una gran cantidad de mano de obra física ya está siendo realizada por máquinas (por ejemplo, fabricación) o pronto será realizada por máquinas (por ejemplo, conducción). Además, una IA suficientemente poderosa será capaz de acelerar el desarrollo de los robots, y luego controlar esos robots en el mundo

físico. Puede ganar algo de tiempo (lo cual es algo bueno), pero me preocupa que no compre mucho. E incluso si la interrupción se limitara solo a tareas cognitivas, seguiría siendo una interrupción sin precedentes grande y rápida.

En tercer lugar, tal vez algunas tareas requieren inherentemente o se benefician enormemente de un toque humano. Estoy un poco más inseguro sobre esto, pero todavía soy escéptico de que sea suficiente para compensar la mayor parte de los impactos que describí anteriormente. La IA ya es ampliamente utilizada para el servicio al cliente. Muchas personas informan que es más fácil hablar con la IA sobre sus problemas personales que hablar con un terapeuta, que la IA es más paciente. Cuando mi hermana estaba luchando con problemas médicos durante un embarazo, sintió que no estaba recibiendo las respuestas o el apoyo que necesitaba de sus proveedores de atención, y encontró que Claude tenía una mejor manera de carretera (así como tener mejor éxito en el diagnóstico del problema). Estoy seguro de que hay algunas tareas para las que un toque humano es realmente importante, pero no estoy seguro de cuántos, y aquí estamos hablando de encontrar trabajo para casi todos en el mercado laboral.

Fourth, some may argue that comparative advantage will still protect humans. Under the law of comparative advantage, even if AI is better than humans at everything, any *relative* differences between the human and AI profile of skills creates a basis of trade and specialization between humans and AI. The problem is that if AIs are literally thousands of times more productive than humans, this logic starts to break down. Even tiny transaction costs could make it not worth it for AI to trade with humans. And human wages may be very low, even if they technically have something to offer.

It's possible all of these factors can be addressed—that the labor market is resilient enough to adapt to even such an enormous disruption. But even if it can eventually adapt, the factors above suggest that the short-term shock will be unprecedented in size.

Defensas

¿Qué podemos hacer con este problema? Tengo varias sugerencias, algunas de las cuales Anthropic ya está haciendo. Lo primero es simplemente obtener datos precisos sobre lo que está sucediendo con el desplazamiento laboral en tiempo real. Cuando un cambio económico ocurre muy rápidamente, es difícil obtener datos confiables sobre lo que está sucediendo, y sin datos confiables es difícil diseñar políticas efectivas. Por ejemplo, los datos del gobierno actualmente carecen de datos granulares de alta frecuencia sobre la adopción de la IA en todas las empresas e industrias. Durante el último año, Anthropic ha estado operando y lanzando públicamente un Índice Económico que muestra el uso de nuestros modelos casi en tiempo real, desglosado por la industria, la tarea, la ubicación e incluso cosas como si una tarea se estaba automatizando o realizando de manera colaborativa. También contamos con un Consejo Asesor Económico para ayudarnos a interpretar estos datos y ver qué viene.

En segundo lugar, las empresas de IA tienen una opción en cómo trabajan con las empresas. La ineficiencia misma de las empresas tradicionales significa que su

despliegue de la IA puede ser muy dependiente del camino, y hay algo de espacio para elegir un mejor camino. Las empresas a menudo tienen la opción de elegir entre "ahorro de costos" (hacer lo mismo con menos personas) y "innovación" (hacer más con el mismo número de personas). El mercado inevitablemente producirá ambos eventualmente, y cualquier compañía competitiva de inteligencia artificial tendrá que servir a algunos de ambos, pero puede haber algo de espacio para dirigir a las empresas hacia la innovación cuando sea posible, y puede ganarnos algo de tiempo. Anthropic está pensando activamente en esto.

En tercer lugar, las empresas deben pensar en cómo cuidar a sus empleados. A corto plazo, ser creativo sobre las formas de reasignar empleados dentro de las empresas puede ser una forma prometedora de evitar la necesidad de despidos. A largo plazo, en un mundo con una enorme riqueza total, en el que muchas empresas aumentan enormemente su valor debido al aumento de la productividad y la concentración de capital, puede ser factible pagar a los empleados humanos incluso mucho tiempo después de que ya no están proporcionando valor económico en el sentido tradicional. Anthropic está considerando actualmente una serie de posibles vías para nuestros propios empleados que compartiremos en un futuro próximo.

En cuarto lugar, las personas ricas tienen la obligación de ayudar a resolver este problema. Es triste para mí que muchas personas ricas (especialmente en la industria tecnológica) hayan adoptado recientemente una actitud cínica y nihilista de que la filantropía es inevitablemente fraudulenta o inútil. Tanto la filantropía privada como la Fundación Gates y programas públicos como PEPFAR han salvado decenas de millones de vidas en el mundo en desarrollo, y han ayudado a crear oportunidades económicas en el mundo desarrollado. Todos los cofundadores de Anthropic se han comprometido a donar el 80% de nuestra riqueza, y el personal de Anthropic se ha comprometido individualmente a donar acciones de la compañía por valor de miles de millones a precios actuales, donaciones que la compañía se ha comprometido a igualar.

En quinto lugar, si bien todas las acciones privadas anteriores pueden ser útiles, en última instancia, un problema macroeconómico tan grande requerirá la intervención del gobierno. La respuesta política natural a un enorme pastel económico junto con una alta desigualdad (debido a la falta de empleos, o empleos mal pagados, para muchos) es la fiscalidad progresiva. El impuesto podría ser general o podría ser dirigido contra las empresas de IA en particular. Obviamente, el diseño fiscal es complicado, y hay muchas maneras de que salga mal. No apoyo políticas fiscales mal diseñadas. Creo que los niveles extremos de desigualdad predichos en este ensayo justifican una política fiscal más robusta por motivos morales básicos, pero también puedo hacer un argumento pragmático a los multimillonarios del mundo de que es de su interés apoyar una buena versión de la misma: si no apoyan una buena versión, inevitablemente obtendrán una mala versión diseñada por una mafia.

En última instancia, pienso en todas las intervenciones anteriores como formas de ganar tiempo. Al final, la IA será capaz de hacer todo, y tenemos que lidiar con eso. Espero que para ese momento, podamos usar la propia IA para ayudarnos a

reestructurar los mercados de manera que funcionen para todos, y que las intervenciones anteriores puedan ayudarnos a superar el período de transición.

Concentración económica del poder

Separado del problema del desplazamiento laboral o la desigualdad económica *per se* está el problema de la *concentración económica del poder.* La Sección 1 discutió el riesgo de que la humanidad sea desempoderada por la IA, y la Sección 3 discutió el riesgo de que los ciudadanos sean desempoderados por sus gobiernos por la fuerza o la coerción. Pero otro tipo de desempoderamiento puede ocurrir si hay una concentración tan grande de riqueza que un pequeño grupo de personas controla efectivamente la política del gobierno con su influencia, y los ciudadanos comunes no tienen influencia porque carecen de influencia económica. La democracia se ve respaldada por la idea de que la población en su conjunto es necesaria para el funcionamiento de la economía. Si esa influencia económica desaparece, entonces el contrato social implícito de la democracia puede dejar de funcionar. Otros han escrito sobre esto, así que no necesito entrar en grandes detalles al respecto aquí, pero estoy de acuerdo con la preocupación, y me preocupa que ya esté empezando a suceder.

Para ser claros, no me opongo a que la gente gane mucho dinero. Hay un fuerte argumento de que incentiva el crecimiento económico en condiciones normales. Simpatizo con las preocupaciones sobre impedir la innovación matando a la gansa de oro que la genera. Pero en un escenario en el que el crecimiento del PIB es del 10 al 20% al año y la IA se está apoderando rápidamente de la economía, sin embargo, los individuos individuales tienen fracciones apreciables del PIB, la innovación *no* es lo que de qué preocuparse. Lo que hay que preocuparse es un nivel de concentración de la riqueza que romperá la sociedad.

El ejemplo más famoso de la concentración extrema de la riqueza en la historia de los Estados Unidos es el Gilded Age, and the wealthiest industrialist of the Gilded Age was John D. Rockefeller. Rockefeller's wealth amounted to ~2% of the US GDP at the time.[42]

[42] La riqueza personal es una "acción", mientras que el PIB es un "flujo", por lo que esto no es una afirmación de que Rockefeller poseía el 2% del valor económico en los Estados Unidos. Pero es más difícil medir la riqueza total de una nación que el PIB, y los ingresos individuales de las personas varían mucho por año, por lo que es difícil hacer una relación en las mismas unidades. La relación entre la mayor fortuna personal y el PIB, aunque no se comparan las manzanas con las manzanas, es, sin embargo, un punto de referencia perfectamente razonable para la concentración extrema de la riqueza.
A similar fraction today would lead to a fortune of $600B, and the richest person in the world today (Elon Musk) already exceeds that, at roughly $700B. So we are already at historically unprecedented levels of wealth concentration, even *before*
La mayor parte del impacto económico de la IA. No creo que sea demasiado exagerado (si conseguimos un "país de genios") imaginar compañías de inteligencia artificial, compañías de semiconductores y tal vez compañías de aplicaciones descendentes que generan ~$3T ingresos por año,[43]

43 El valor total de la mano de obra en toda la economía es de $ 60T / año, por lo que $ 3T / año correspondería al 5% de esto. Esa cantidad podría ser ganada por una compañía que suministró mano de obra por el 20% del costo de los humanos y tenía una participación de mercado del 25%, incluso si la demanda de mano de obra no se expandiera (lo que casi con seguridad sería debido al menor costo).

Ser valorado en ~$30T, y llevar a fortunas personales hasta bien entrados los billones. En ese mundo, los debates que tenemos sobre la política fiscal de hoy simplemente no se aplicarán, ya que estaremos en una situación fundamentalmente diferente.

En relación con esto, el acoplamiento de esta concentración económica de la riqueza con el sistema político ya me preocupa. Los centros de datos de IA ya representan una fracción sustancial del crecimiento económico de los Estados Unidos.[44]

[44] Para ser claros, no creo que la productividad real de la IA sea aún responsable de una fracción sustancial del crecimiento económico de los Estados Unidos. Más bien, creo que el gasto en el centro de datos representa el crecimiento causado por la inversión anticipada que equivale al mercado que espera un *futuro* crecimiento económico impulsado por la IA e invierte en consecuencia.
and are thus strongly tying together the financial interests of large tech companies (which are increasingly focused on either AI or AI infrastructure) and the political interests of the government in a way that can produce perverse incentives. We already see this through the reluctance of tech companies to criticize the US government, and the government's support for extreme anti-regulatory policies on AI.

Defenses

What can be done about this? First, and most obviously, companies should simply choose not to be part of it. Anthropic has always strived to be a policy actor and not a political one, and to maintain our authentic views whatever the administration. We've spoken up in favor of Regulación sensible de la IA and export controls that are in the public interest, even when these are at odds with government policy.[45]

[45] Cuando estamos de acuerdo con la administración, lo decimos, y buscamos puntos de acuerdo donde las políticas apoyadas mutuamente sean realmente buenas para el mundo. Nuestro objetivo es ser corredores honestos en lugar de partidarios u opositores de un partido político dado.
Many people have told me that we should stop doing this, that it could lead to unfavorable treatment, but in the year we've been doing it, Anthropic's valuation has increased by over 6x, an almost unprecedented jump at our commercial scale.

En segundo lugar, la industria de la IA necesita una relación más saludable con el gobierno, una basada en el compromiso político sustantivo en lugar de la alineación política. Nuestra elección de participar en la sustancia política en lugar de la política a veces se lee como un error táctico o la falta de "leer la habitación" en lugar de una decisión de principios, y ese marco me preocupa. En una democracia saludable, las empresas deberían poder abogar por una buena política por su propio bien. En relación con esto, se está elaborando una reacción pública

contra la IA: esto podría ser un correctivo, pero 'actualmente no está enfocado. Gran parte de ella se dirige a problemas que en realidad no son problemas (como el uso de agua del centro de datos) y propone soluciones (como prohibiciones de centros de datos o impuestos sobre la riqueza mal diseñados) que no abordarían las preocupaciones reales. El problema subyacente que merece atención es garantizar que el desarrollo de la IA siga siendo responsable del interés público, no capturado por ninguna alianza política o comercial en particular, y parece importante centrar la discusión pública allí.

En tercer lugar, las intervenciones macroeconómicas que describí anteriormente en esta sección, así como un resurgimiento de la filantropía privada, pueden ayudar a equilibrar las escalas económicas, abordando tanto el desplazamiento laboral como la concentración de los problemas de poder económico a la vez. Deberíamos mirar la historia de nuestro país aquí: incluso en la Edad Dorada, industriales como Rockefeller y Carnegie sintieron una fuerte obligación con la sociedad en general, un sentimiento de que la sociedad había contribuido enormemente a su éxito y que necesitaban retribuir. Ese espíritu parece faltar cada vez más hoy, y creo que es una gran parte de la salida de este dilema económico. Aquellos que están a la vanguardia del auge económico de la IA deberían estar dispuestos a regalar su riqueza y su poder.

5. Mares negros del infinito

Efectos indirectos

This last section is a catchall for unknown unknowns, particularly things that could go wrong as an indirect result of positive advances in AI and the resulting acceleration of science and technology in general. Suppose we address all the risks described so far, and begin to reap the benefits of AI. We will likely get a "century of scientific and economic progress compressed into a decade," and this will be hugely positive for the world, but we will then have to contend with the problems that arise from this rapid rate of progress, and those problems may come at us fast. We may also encounter other risks that occur indirectly as a consequence of AI progress and are hard to anticipate in advance.

Por la naturaleza de incógnitas desconocidas es imposible hacer una lista exhaustiva, pero enumeraré tres posibles preocupaciones como ejemplos ilustrativos de lo que deberíamos estar esperando:

- **Rápidos avances en biología. Si obtenemos un siglo de progreso médico en unos pocos años, es posible que aumentemos enormemente la vida humana, y existe la posibilidad de que también obtengamos capacidades radicales como la capacidad de aumentar la inteligencia humana o modificar radicalmente la biología humana. Esos serían grandes cambios en lo que es posible, sucediendo muy rápidamente. Podrían ser positivos si se hacen responsablemente (que es mi esperanza, como se describe en *Máquinas de la Gracia Amorosa*), pero siempre existe el riesgo de que vayan muy mal, por ejemplo, si los esfuerzos para hacer que los humanos sean más inteligentes también los hacen más inestables o en busca de energía. También está el tema de las uploads"cargas" o "emulación de todo el cerebro", las mentes**

humanas digitales instanciadas en el software, que algún día podrían ayudar a la humanidad a trascender sus limitaciones físicas, pero que también conllevan riesgos que encuentro inquietantes.

- **La IA cambia la vida humana de una manera poco saludable. Un mundo con miles de millones de inteligencias que son mucho más inteligentes que los humanos en todo va a ser un mundo muy extraño para vivir. Incluso si la IA no tiene como objetivo atacar activamente a los humanos (Sección 1), y no se usa explícitamente para la opresión o el control por parte de los estados (Sección 3), hay mucho que podría salir mal antes de esto, a través de incentivos comerciales normales y transacciones nominalmente consensuadas. Vemos los primeros indicios de esto en las preocupaciones sobre la psicosis de la IA, la IA que lleva a la gente al suicidio y las preocupaciones sobre las relaciones románticas con las IA. Como ejemplo, ¿podrían las poderosas IA inventar alguna nueva religión y convertir a millones de personas en ella? ¿Podría la mayoría de la gente terminar "adicto" de alguna manera a las interacciones de IA? ¿Podría la gente terminar siendo "títere" por los sistemas de inteligencia artificial, donde una IA esencialmente observa cada uno de sus movimientos y les dice exactamente qué hacer y decir en todo momento, lo que lleva a una vida "buena" pero que carece de libertad o de algún orgullo de logro? No sería difícil generar docenas de estos escenarios si me sentara con el creador de *Black Mirror* y tratara de hacer una lluvia de ideas. Creo que esto apunta a la importancia de cosas como mejorar la Constitución de Claude, más allá de lo necesario para prevenir los problemas en la Sección 1. Asegurarse de que los modelos de IA *realmente* tengan los intereses a largo plazo de sus usuarios en el corazón, de una manera que las personas reflexivas respaldarían en lugar de alguna manera sutilmente distorsionada, parece crítico.**
- **Human purpose. This is related to the previous point, but it's not so much about specific human interactions with AI systems as it is about how human life changes in general in a world with powerful AI. Will humans be able to find purpose and meaning in such a world? I think this is a matter of attitude: as I said in *Machines of Loving Grace*, I think human purpose does not depend on being the best in the world at something, and humans can find purpose even over very long periods of time through stories and projects that they love. We simply need to break the link between the generation of economic value and self-worth and meaning. But that is a transition society has to make, and there is always the risk we don't handle it well.**

Mi esperanza con todos estos problemas potenciales es que en un mundo con una IA poderosa en el que confiamos para no matarnos, que no es la herramienta de un gobierno opresivo, y que está trabajando genuinamente en nuestro nombre, podemos usar la propia IA para anticipar y prevenir estos problemas. Pero eso no está garantizado, como todos los otros riesgos, es algo que tenemos que manejar con cuidado.

La prueba de la humanidad

Leer este ensayo puede dar la impresión de que estamos en una situación desalentadora. Ciertamente me pareció desalentador escribir, en contraste *con Machines of Loving Grace*, que tenía ganas de dar forma y estructura a la música extremadamente hermosa que había estado haciendo eco en mi cabeza durante años. Y hay mucho sobre la situación que realmente es difícil. La IA trae amenazas a la humanidad desde múltiples direcciones, y hay una tensión genuina entre los diferentes peligros, donde mitigar algunos de ellos corre el riesgo de empeorar a otros si no enhebramos la aguja con mucho cuidado.

Tomarse el tiempo para construir cuidadosamente sistemas de inteligencia artificial para que no amenacen de manera autónoma a la humanidad está en tensión genuina con la necesidad de que las naciones democráticas se mantengan por delante de las naciones autoritarias y no sean subyugadas por ellas. Pero a su vez, las mismas herramientas habilitadas para la IA que son necesarias para combatir las autocracias pueden, si se llevan demasiado lejos, volverse hacia adentro para crear tiranía en nuestros propios países. El terrorismo impulsado por la inteligencia artificial podría matar a millones de personas a través del uso indebido de la biología, pero una reacción exagerada a este riesgo podría llevarnos por el camino hacia un estado de vigilancia autocrático. Los efectos laborales y de concentración económica de la IA, además de ser graves problemas por derecho propio, pueden obligarnos a enfrentar los otros problemas en un ambiente de ira pública y tal vez incluso disturbios civiles, en lugar de poder recurrir a los mejores ángeles de nuestra naturaleza. Sobre todo, la gran *cantidad* de riesgos, incluidos los desconocidos, y la necesidad de lidiar con todos ellos a la vez, crea un guante intimidante que la humanidad debe correr.

Además, los últimos años deberían dejar claro que la idea de detener o incluso ralentizar sustancialmente la tecnología es fundamentalmente insostenible. La fórmula para construir potentes sistemas de inteligencia artificial es increíblemente simple, tanto que casi se puede decir que emerge espontáneamente de la combinación correcta de datos y computación en bruto. Su creación fue probablemente inevitable en el instante en que la humanidad inventó el transistor, o posiblemente incluso antes cuando aprendimos por primera vez a controlar el fuego. Si una empresa no lo construye, otros lo harán casi tan rápido. Si todas las empresas de los países democráticos detuvieran o ralentizaran el desarrollo, de mutuo acuerdo o decreto regulatorio, entonces los países autoritarios simplemente seguirían adelante. Dado el increíble valor económico y militar de la tecnología, junto con la falta de un mecanismo de aplicación significativo, no veo cómo podríamos convencerlos de que se detengan.

I do see a path to a *slight* moderation in AI development that is compatible with a realist view of geopolitics. That path involves slowing down the march of autocracies towards powerful AI for a few years by denying them the resources they need to build it,[46]

[46] No creo que sea posible nada más que unos pocos años: en escalas de tiempo más largas, construirán sus propios chips.
namely chips and semiconductor manufacturing equipment. This in turn gives democratic countries a buffer that they can "spend" on building powerful AI more carefully, with more attention to its risks, while still proceeding fast enough to

comfortably beat the autocracies. The race between AI companies within democracies can then be handled under the umbrella of a common legal framework, via a mixture of industry standards and regulation.

Anthropic ha abogado muy duro por este camino, al presionar por los controles de exportación de chips y la regulación juiciosa de la inteligencia artificial, pero incluso estas propuestas aparentemente de sentido común han sido rechazadas en gran medida por los responsables políticos en los Estados Unidos (que es el país donde es más importante tenerlas). Hay tanto dinero que se puede ganar con la IA, literalmente billones de dólares por año, que incluso las medidas más simples están teniendo dificultades para superar la economía política inherente a la IA. Esta es la trampa: la IA es tan poderosa, un premio tan brillante, que es muy difícil para la civilización humana imponerle restricciones.

Puedo imaginar, como lo hizo Sagan en *Contact,* que esta misma historia se desarrolla en miles de mundos. Una especie gana sensibilidad, aprende a usar herramientas, comienza el ascenso exponencial de la tecnología, se enfrenta a las crisis de la industrialización y las armas nucleares, y si sobrevive a ellas, se enfrenta al desafío más difícil y final cuando aprende a dar forma a la arena en máquinas que piensan. Si sobrevivimos a esa prueba y continuamos construyendo la hermosa sociedad descrita en *Máquinas de la Gracia Amorosa,* o sucumbimos a la esclavitud y la destrucción, dependerá de nuestro carácter y nuestra determinación como especie, nuestro espíritu y nuestra alma.

A pesar de los muchos obstáculos, creo que la humanidad tiene la fuerza en su interior para pasar esta prueba. Me siento alentado e inspirado por los miles de investigadores que han dedicado sus carreras a ayudarnos a comprender y dirigir los modelos de IA, y a dar forma al carácter y la constitución de estos modelos. Creo que ahora hay una buena posibilidad de que esos esfuerzos fructifiquen a tiempo para la materia. Me alienta que al menos algunas compañías hayan declarado que '"pagarán costos comerciales significativos para impedir que sus modelos contribuyan a la amenaza del bioterrorismo. Me alienta que algunas personas valientes se hayan resistido a los vientos políticos prevalecientes y aprobado una legislación que pone las primeras semillas de barandillas sensatas en los sistemas de inteligencia artificial. Me alienta que el público entienda que la IA conlleva riesgos y quiere que se aborden esos riesgos. Me siento alentado por el indomable espíritu de libertad en todo el mundo y la determinación de resistir la tiranía dondequiera que ocurra.

Pero tendremos que intensificar nuestros esfuerzos si queremos tener éxito. El primer paso es que los más cercanos a la tecnología simplemente digan la verdad sobre la situación en la que está la humanidad, lo que siempre he tratado de hacer; lo estoy haciendo más explícitamente y con mayor urgencia con este ensayo. El siguiente paso será convencer a los pensadores, políticos, empresas y ciudadanos del mundo de la inminencia y la importancia primordial de este tema, de que vale la pena gastar pensamiento y capital político en esto en comparación con los miles de otros temas que dominan las noticias todos los días. Entonces habrá un tiempo para el coraje, para que suficientes personas se resistan a las tendencias prevalecientes y se mantengan en principio, incluso frente a las amenazas a sus intereses económicos y la seguridad personal.

Los años frente a nosotros serán imposiblemente difíciles, preguntándonos más de lo que creemos que podemos dar. Pero en mi tiempo como investigador, líder y ciudadano, he visto suficiente coraje y nobleza para creer que podemos ganar, que cuando se pone en las circunstancias más oscuras, la humanidad tiene una forma de reunirse, aparentemente en el último minuto, la fuerza y la sabiduría necesarias para prevalecer. No tenemos tiempo que perder.

Me gustaría agradecer a Erik Brynjolfsson, Ben Buchanan, Mariano-Florentino Cuéllar, Allan Dafoe, Kevin Esvelt, Nick Beckstead, Richard Fontaine, Jim McClave y muchos de los empleados de Anthropic por sus útiles comentarios sobre los borradores de este ensayo.

Notas a pie de página

1. [1] **Esto es simétrico hasta un punto que hice en *Máquinas de Gracia Amorosa*, donde comencé diciendo que las ventajas de la IA no deberían pensarse en términos de una profecía de salvación, y que es importante ser concreto y fundamentado y evitar la grandiosidad. En última instancia, las profecías de salvación y las profecías de la perdición no son útiles para enfrentar el mundo real, básicamente por las mismas razones.** ↩
2. [2] **Anthropic's goal is to remain consistent through such changes. When talking about AI risks was politically popular, Anthropic cautiously advocated for a judicious and evidence-based approach to these risks. Now that talking about AI risks is politically unpopular, Anthropic continues to cautiously advocate for a judicious and evidence-based approach to these risks.**↩
3. [3] **Con el tiempo, he ganado una confianza cada vez mayor en la trayectoria de la IA y la probabilidad de que supere la capacidad humana en todos los ámbitos, pero aún persiste cierta incertidumbre.** ↩
4. [4] **Los controles de exportación de chips son un gran ejemplo de esto. Son simples y parecen funcionar principalmente.** ↩
5. [5] **And of course, the hunt for such evidence must be intellectually honest, such that it could also turn up evidence of a lack of danger. Transparency through model cards and other disclosures is an attempt at such an intellectually honest endeavor.**↩
6. [6] **De hecho, desde que se escribió *Machines of Loving Grace* en 2024, los sistemas de inteligencia artificial se han vuelto capaces de hacer tareas que requieren varias horas, y METR recientemente evaluó que el Opus 4.5 puede hacer aproximadamente cuatro horas de trabajo humanas con un 50% de confiabilidad.** ↩
7. [7] **Y para ser claros, incluso si la poderosa IA está a solo 1-2 años de distancia en un sentido técnico, muchas de sus consecuencias sociales, tanto positivas como negativas, pueden tardar unos años más en ocurrir. Esta es la razón por la que puedo pensar simultáneamente que la IA interrumpirá el 50% de los trabajos de cuello blanco de *nivel* de *entrada* durante 1 a 5 años, mientras que también creo que podemos tener una IA que sea más capaz que *todos* en solo 1-2 años.** ↩

8. [8] **Vale la pena agregar que el *público* (en comparación con los responsables políticos) parece estar muy preocupado por los riesgos de la IA. Creo que parte de su enfoque es correcto (es decir, El desplazamiento del trabajo de la IA), y algunos son equivocados (como las preocupaciones sobre el uso del agua de la IA, que no es significativo). Esta reacción me da esperanza de que es posible un consenso sobre cómo abordar los riesgos, pero hasta ahora aún no se ha traducido en cambios de política, y mucho menos en cambios de política efectivos o bien orientados.** ↩
9. [9] **También pueden, por supuesto, manipular (o simplemente pagar) a un gran número de humanos para que hagan lo que quieren en el mundo físico.** ↩
10. [10] **No creo que este sea un hombre de paja: es mi entendimiento, por ejemplo, que Yann LeCun mantiene esta posición.** ↩
11. [11] **Por ejemplo, véase la Sección 5.5.2 (p. 63–66) de la tarjeta del sistema Claude 4.** ↩
12. [12] **También hay una serie de otras suposiciones inherentes al modelo simple, que no voy a discutir aquí. En términos generales, deberían hacernos menos preocupados por la simple historia específica de la búsqueda de poder desalineada, pero también más preocupados por el posible comportamiento impredecible que no hemos anticipado.** ↩
13. [13] ***Ender's Game*** **describes a version of this involving humans rather than AI.**↩
14. [14] **For example, models may be told not to do various bad things, and also to obey humans, but may then observe that many humans do exactly those bad things! It's not clear how this contradiction would resolve (and a well-designed constitution should encourage the model to handle these contradictions gracefully), but this type of dilemma is not so different from the supposedly "artificial" situations that we put AI models in during testing.**↩
15. [15] **Por cierto, una consecuencia de que la constitución sea un documento de lenguaje natural es que es legible para el mundo, y eso significa que puede ser criticado por cualquier persona y comparado con documentos similares por otras empresas. Sería valioso crear una carrera hacia la cima que no solo anime a las empresas a publicar estos documentos, sino que los anime a ser buenos.** ↩
16. [16] **Incluso hay una hipótesis sobre un principio unificador profundo que conecta el enfoque basado en el carácter de la IA constitucional para obtener resultados de la capacidad de interpretación y la ciencia de la alineación. Según la hipótesis, los mecanismos fundamentales que impulsan a Claude surgieron originalmente como formas de simular personajes en el preentrenamiento, como predecir lo que dirían los personajes de una novela. Esto sugeriría que una forma útil de pensar acerca de la constitución es más como una descripción de carácter que el modelo utiliza para instanciar una persona consistente. También nos ayudaría a explicar los resultados de "debo ser una mala persona" que mencioné anteriormente (porque el modelo está tratando de *actuar como si* fuera un carácter coherente, en este caso uno malo), y sugeriría que los métodos de interpretabilidad deberían poder descubrir "rasgos psicológicos" dentro de los modelos. Nuestros investigadores están trabajando en formas de probar esta hipótesis.** ↩

17. [17] **Para ser claros, el monitoreo se realiza de manera que preserva la privacidad.** ↩
18. [18] **Incluso en nuestros propios experimentos con lo que esencialmente se imponen reglas voluntarias con nuestra Política de Escalamiento Responsable, hemos encontrado una y otra vez que es muy fácil terminar siendo demasiado rígido, dibujando líneas que parecen importantes ex ante pero que resultan ser tontas en retrospectiva. Es muy fácil establecer reglas sobre las cosas equivocadas cuando una tecnología avanza rápidamente.** ↩
19. [19] **SB 53 y RAISE no se aplican en absoluto a empresas con menos de $ 500M en ingresos anuales. Solo se aplican a empresas más grandes y establecidas como Anthropic.** ↩
20. [20] **Originalmente leí el ensayo de Joy hace 25 años, cuando fue escrito, y tuvo un profundo impacto en mí. Entonces y ahora, lo veo como demasiado pesimista, no creo que la "renuncia" amplia de áreas enteras de la tecnología, lo que Joy sugiere, sea la respuesta, pero los problemas que plantea fueron sorprendentemente proféticos, y Joy también escribe con un profundo sentido de compasión y humanidad que admiro.** ↩
21. [21] **Tenemos que preocuparnos por los actores estatales, ahora y en el futuro, y lo discuto en la siguiente sección.** ↩
22. [22] **There is evidence that many terrorists are at least relatively well-educated, which might seem to contradict what I'm arguing here about a negative correlation between ability and motivation. But I think in actual fact they are compatible observations: if the ability threshold for a successful attack is high, then almost by definition those who *currently* succeed must have high ability, even if ability and motivation are negatively correlated. But in a world where the limitations on ability were removed (e.g., with future LLMs), I'd predict that a substantial population of people with the motivation to kill but lower ability would start to do so—just as we see for crimes that don't require much ability (like school shootings).**↩
23. [23] **Sin embargo, Aum Shinrikyo lo intentó. El líder de Aum Shinrikyo, Seiichi Endo, se formó en virología de la Universidad de Kyoto e intentó producir ántrax y ébola. Sin embargo, a partir de 1995, incluso carecía de suficiente experiencia y recursos para tener éxito en esto. La barra es ahora sustancialmente más baja, y los LLM podrían reducirla aún más.** ↩
24. [24] **A bizarre phenomenon relating to mass murderers is that the style of murder they choose operates almost as a grotesque sort of fad. In the 1970s and 1980s, serial killers were very common, and new serial killers often copied the behavior of more established or famous serial killers. In the 1990s and 2000s, mass shootings became more common, while serial killers became less common. There is no technological change that triggered these patterns of behavior, it just appears that violent murderers were copying each others' behavior and the "popular" thing to copy changed.**↩
25. [25] **Los despilfarradores casuales a veces creen que han comprometido estos clasificadores cuando consiguen que el modelo produzca una pieza específica de información, como la secuencia del genoma de un virus. Pero como expliqué antes, el modelo de amenaza que nos preocupa implica un consejo interactivo paso a paso que se extiende a lo largo de semanas o meses sobre pasos oscuros específicos en el proceso de producción de armas biológicas, y esto es contra lo que nuestros clasificadores pretenden defenderse. (A menudo describimos nuestra investigación como la búsqueda**

de jailbreaks “universales”, que no solo funcionan en un contexto específico o estrecho, sino que abren ampliamente el comportamiento del modelo). ↩

26. **[26] Though we will continue to invest in work to make our classifiers more efficient, and it may make sense for companies to share advances like these with one another.↩**
27. **[27] Obviously, I do not think companies should have to disclose technical details about the specific steps in biological weapons production that they are blocking, and the transparency legislation that has been passed so far (SB 53 and RAISE) accounts for this issue.↩**
28. **[28] Another related idea is “resilience markets” where the government encourages stockpiling of PPE, respirators, and other essential equipment needed to respond to a biological attack by promising ahead of time to pay a pre-agreed price for this equipment in an emergency. This incentivizes suppliers to stockpile such equipment without fear that the government will seize it without compensation.↩**
29. **[29] Why am I more worried about large actors for seizing power, but small actors for causing destruction? Because the dynamics are different. Seizing power is about whether one actor can amass enough strength to overcome everyone else—thus we should worry about the most powerful actors and/or those closest to AI. Destruction, by contrast, can be wrought by those with little power if it is much harder to defend against than to cause. It is then a game of defending against the most *numerous* threats, which are likely to be smaller actors.↩**
30. **[30] This might sound like it is in tension with my point that attack and defense may be more balanced with cyberattacks than with bioweapons, but my worry here is that if a country’s AI is the most powerful in the world, then others will not be able to defend even if the technology itself has an intrinsic attack-defense balance.↩**
31. **[31] For example, in the United States this includes the fourth amendment and the Posse Comitatus Act.↩**
32. **[32] Además, para ser claros, hay algunos argumentos para construir grandes centros de datos en países con diferentes estructuras de gobernanza, especialmente si están controlados por empresas en democracias. Tales construcciones podrían, en principio, ayudar a las democracias a competir mejor con el PCCh, que es la mayor amenaza. También creo que estos centros de datos no representan mucho riesgo a menos que sean muy grandes. Pero en general, creo que se justifica la precaución al colocar centros de datos muy grandes en países donde las salvaguardias institucionales y las protecciones del estado de derecho están menos bien establecidas. ↩**
33. **[33] Este es, por supuesto, también un argumento para mejorar la seguridad de la disuasión nuclear para que sea más probable que sea robusta contra la poderosa IA, y las democracias con armas nucleares deberían hacerlo. Pero no sabemos de qué será capaz una poderosa IA o qué defensas, si las hay, funcionarán en contra, por lo que no debemos asumir que estas medidas necesariamente resolverán el problema. ↩**
34. **[34] También existe el riesgo de que, incluso si la disuasión nuclear sigue siendo efectiva, un país atacante decida llamar a nuestro farol, no está claro si estaríamos dispuestos a usar armas nucleares para defendernos contra un enjambre de aviones no tripulados, incluso si el enjambre de aviones no**

tripulados tiene un riesgo sustancial de conquistarnos. Los enjambres de drones pueden ser algo nuevo que es menos severo que los ataques nucleares, pero más severo que los ataques convencionales. Alternativamente, diferentes evaluaciones de la efectividad de la disuasión nuclear en la era de la IA podrían alterar la teoría de juegos del conflicto nuclear de una manera desestabilizadora. ↩

35. **[35] Para ser claros, creo que es la estrategia correcta no vender chips a China, incluso si la línea de tiempo para una IA poderosa fuera sustancialmente más larga. No podemos hacer que los chinos sean "adictos" a los chips estadounidenses, están decididos a desarrollar su industria de chips nativos de una manera u otra. Les llevará muchos años hacerlo, y todo lo que estamos haciendo vendiendo esos chips les está dando un gran impulso durante ese tiempo.** ↩
36. **[36] Para ser claros, la mayor parte de lo que se está utilizando en Ucrania y Taiwán hoy en día no son armas *totalmente* autónomas. Estos vienen, pero no aquí hoy.** ↩
37. **[37] Nuestra tarjeta modelo para Claude Opus 4.5, nuestro modelo más reciente, muestra que Opus tiene un mejor rendimiento en una entrevista de ingeniería de rendimiento frecuentemente dada en Anthropic que cualquier entrevistado en la historia de la empresa.** ↩
38. **[38] "Escribir todo el código" y "hacer la tarea de un ingeniero de software de extremo a extremo" son cosas muy diferentes, porque los ingenieros de software hacen mucho más que escribir código, incluyendo pruebas, tratar con entornos, archivos e instalación, administrar implementaciones de computación en la nube, iterar en productos y mucho más.** ↩
39. **[39] Las computadoras son generales en cierto sentido, pero son claramente incapaces por sí solas de la gran mayoría de las habilidades cognitivas humanas, incluso cuando exceden en gran medida a los humanos en unas pocas áreas (como la aritmética). Por supuesto, las cosas construidas *sobre* las computadoras, como la inteligencia artificial, ahora son capaces de una amplia gama de habilidades cognitivas, que es de lo que se trata este ensayo.** ↩
40. **[40] Para ser claros, los modelos de IA no tienen precisamente el mismo perfil de fortalezas y debilidades que los humanos. Pero también están avanzando de manera bastante uniforme a lo largo de cada dimensión, de tal manera que tener un perfil puntiagudo o desigual puede no importar en última instancia.** ↩
41. **[41] Aunque hay debate entre los economistas sobre esta idea.** ↩
42. **[42] La riqueza personal es una "acción", mientras que el PIB es un "flujo", por lo que esto no es una afirmación de que Rockefeller poseía el 2% del valor económico en los Estados Unidos. Pero es más difícil medir la riqueza total de una nación que el PIB, y los ingresos individuales de las personas varían mucho por año, por lo que es difícil hacer una relación en las mismas unidades. La relación entre la mayor fortuna personal y el PIB, aunque no se comparan las manzanas con las manzanas, es, sin embargo, un punto de referencia perfectamente razonable para la concentración extrema de la riqueza.** ↩
43. **[43] El valor total de la mano de obra en toda la economía es de $ 60T / año, por lo que $ 3T / año correspondería al 5% de esto. Esa cantidad podría ser ganada por una compañía que suministró mano de obra por el 20% del**

costo de los humanos y tenía una participación de mercado del 25%, incluso si la demanda de mano de obra no se expandiera (lo que casi con seguridad sería debido al menor costo). ↩

44. **[44] Para ser claros, no creo que la productividad real de la IA sea aún responsable de una fracción sustancial del crecimiento económico de los Estados Unidos. Más bien, creo que el gasto en el centro de datos representa el crecimiento causado por la inversión anticipada que equivale al mercado que espera un *futuro* crecimiento económico impulsado por la IA e invierte en consecuencia.** ↩
45. **[45] Cuando estamos de acuerdo con la administración, lo decimos, y buscamos puntos de acuerdo donde las políticas apoyadas mutuamente sean realmente buenas para el mundo. Nuestro objetivo es ser corredores honestos en lugar de partidarios u opositores de un partido político dado.** ↩
46. **[46] No creo que sea posible nada más que unos pocos años: en escalas de tiempo más largas, construirán sus propios chips.** ↩