

Leopold Aschenbrenner

SITUATIONAL AWARENESS

La década que se avecina

JUNE 2024

Dedicado a Ilya Sutskever.

Si bien solía trabajar en OpenAI, todo esto se basa en información disponible públicamente, mis propias ideas, conocimiento general del campo o chismes de ciencia ficción.

Gracias a Collin Burns, Avital Balwit, Carl Shulman, Jan Leike, Ilya Sutskever, Holden Karnofsky, Sholto Douglas, James Bradbury, Dwarkesh Patel y muchos otros por las discusiones formativas. Gracias a muchos amigos por sus comentarios sobre los borradores anteriores. Gracias a Joe Ronan por su ayuda con los gráficos, y Nick Whitaker por su ayuda con la publicación.

situational-awareness.ai
leopold@SITUATIONAL-
AWARENESS.AI Actualizado el 6 de
junio de 2024
San Francisco, California

Puedes ver el futuro primero en San Francisco.

En el último año, la comidilla de la ciudad ha pasado de clústeres de cómputo de 10.000 millones de dólares a clústeres de 100.000 millones de dólares y clústeres de un billón de dólares. Cada seis meses se añade otro cero a los planes de la sala de juntas. Detrás de escena, hay una lucha feroz para asegurar todos los contratos de energía que aún estén disponibles para el resto de la década, todos los transformadores de voltaje que se puedan adquirir. Las grandes empresas estadounidenses se están preparando para invertir billones de dólares en una movilización del poderío industrial estadounidense, que no se ha visto desde hace mucho tiempo. A finales de la década, la producción de electricidad estadounidense habrá crecido decenas por ciento; desde los campos de esquisto de Pensilvania hasta las granjas solares de Nevada, cientos de millones de GPU zumbarán.

La carrera de AGI ha comenzado. Estamos construyendo máquinas que pueden pensar y razonar. Para 2025/26, estas máquinas superarán a los graduados universitarios. Al final de la década, serán más inteligentes que tú o que yo; Tendremos superinteligencia, en el verdadero sentido de la palabra. En el camino, las fuerzas de seguridad nacional que no se habían visto en medio siglo se desatarán y, en poco tiempo, El Proyecto estará en marcha. Si tenemos suerte, estaremos en una carrera sin cuartel con el PCCh; Si tenemos mala suerte, una guerra sin cuartel.

Todo el mundo habla ahora de la IA, pero pocos tienen el más mínimo atisbo de lo que está a punto de golpearlos. Los analistas de Nvidia siguen pensando que 2024 podría estar cerca del pico. Los expertos de la corriente principal están atrapados en la ceguera voluntaria de "solo se trata de predecir la siguiente palabra". Lo único que ven es exageración y negocios como de costumbre; A lo sumo, contemplan otro cambio tecnológico a escala de Internet.

En poco tiempo, el mundo despertará. Pero en este momento, hay quizás unos pocos cientos de personas, la mayoría de ellas en San Francisco y en los laboratorios de IA, que tienen *conciencia situacional*. A través de las peculiares fuerzas del destino, me he encontrado entre ellos. Hace unos años, estas personas eran ridiculizadas como locas, pero confiaban en las líneas de tendencia, lo que les permitía predecir correctamente los avances de la IA de los últimos años. Queda por ver si estas personas también tienen razón sobre los próximos años. Pero estas son personas muy inteligentes, las personas más inteligentes que he conocido, y son las que están construyendo esta tecnología. Tal vez sean una extraña nota a pie de página en la historia, o tal vez pasen a la historia como Szilard, Oppenheimer y Teller. Si están viendo el futuro aunque sea de cerca, nos espera un viaje salvaje.

Déjame decirte lo que vemos.

Contenido

Introducción 3

La historia está viva en San Francisco.

I. De GPT-4 a AGI: Conteo de los OOM 7

El AGI para 2027 es sorprendentemente plausible. GPT-2 a GPT-4 nos llevó de ~preescolar a ~habilidades inteligentes de estudiante de secundaria en 4 años. Seguimiento de tendencias-líneas en el proceso (~0.5 órdenes de magnitud o OOMs/año), eficiencias algorítmicas (~0.5 OOMs/año), y las ganancias "descojeando" (de chat-bot a agente), deberíamos esperar otro salto cualitativo del tamaño de un niño en edad preescolar a un niño de secundaria para 2027.

II. De la AGI a la superinteligencia: la explosión de la inteligencia 46

El progreso de la IA no se detendrá a nivel humano. Cientos de millones de AGI podrían automatizar la investigación de IA, comprimiendo una década de progreso algorítmico (5+ OOMs) en 1 año. Pasaríamos rápidamente del nivel humano a sistemas de IA enormemente sobrehumanos. El poder —y el peligro— de la superinteligencia sería dramático.

III. Los desafíos 74

IIIa. Corriendo hacia el grupo del billón de dólares 75

La aceleración más extraordinaria del tecnocapital se ha puesto en marcha. A medida que los ingresos de la IA crezcan rápidamente, se destinarán muchos billones de dólares a la construcción de GPU, centros de datos y energía antes del final de la década. La movilización industrial, incluido el aumento de la producción de electricidad en Estados Unidos en un 10 por ciento, será intensa.

IIIb. Bloqueo de los laboratorios: seguridad para AGI 89

Los principales laboratorios de IA del país tratan la seguridad como una ocurrencia tardía. Actualmente, básicamente están entregando los secretos clave de la AGI al PCCh en bandeja de plata. Asegurar los secretos y pesos de AGI contra la amenaza de los actores estatales será un esfuerzo inmenso, y no vamos por buen camino.

IIIc. Superalineación **105**

Controlar de forma fiable sistemas de IA mucho más inteligentes que nosotros es un problema técnico sin resolver. Y aunque es un problema solucionable, las cosas podrían descarrilarse muy fácilmente durante una rápida explosión de inteligencia. Manejar esto será extremadamente tenso; El fracaso podría ser fácilmente catastrófico.

IIIId. El mundo libre debe prevalecer **126**

La superinteligencia dará una ventaja económica y militar decisiva. China aún no está fuera de juego. En la carrera hacia el AGI, la supervivencia misma del mundo libre estará en juego. ¿Podemos mantener nuestra preeminencia sobre los poderes autoritarios? ¿Y lograremos evitar la autodestrucción en el camino?

IV. El proyecto **141**

A medida que se intensifique la carrera hacia la AGI, el estado de seguridad nacional se verá involucrado. El Gobierno de los Estados Unidos despertará de su letargo, y para el 27/28 tendremos algún tipo de proyecto AGI del gobierno. Ninguna startup puede manejar la superinteligencia. En algún lugar de un SCIF, el final del juego estará en marcha.

V. Pensamientos de despedida **156**

¿Y si estamos en lo cierto?

Apéndice **162**

I. De GPT-4 a AGI: Contando los OOM

El AGI para 2027 es sorprendentemente plausible. GPT-2 a GPT-4 nos llevó de ~preescolares a ~inteligentes habilidades de estudiante de secundaria en 4 años. Seguimiento de líneas de tendencia en el proceso (~0.5 órdenes de magnitud o OOMs/año), eficiencias algorítmicas (~0.5 OOMs/año), y las ganancias "sin cojear" (de chatbot a agente), deberíamos esperar otro salto cualitativo del tamaño de un niño en edad preescolar a un niño de secundaria para 2027.

Mirar. Los modelos, solo quieren aprender. Tienes que entender esto. Los modelos, solo quieren aprender.

ILYA Sutskever
(circa 2015, vía [Dario Amodei](#))

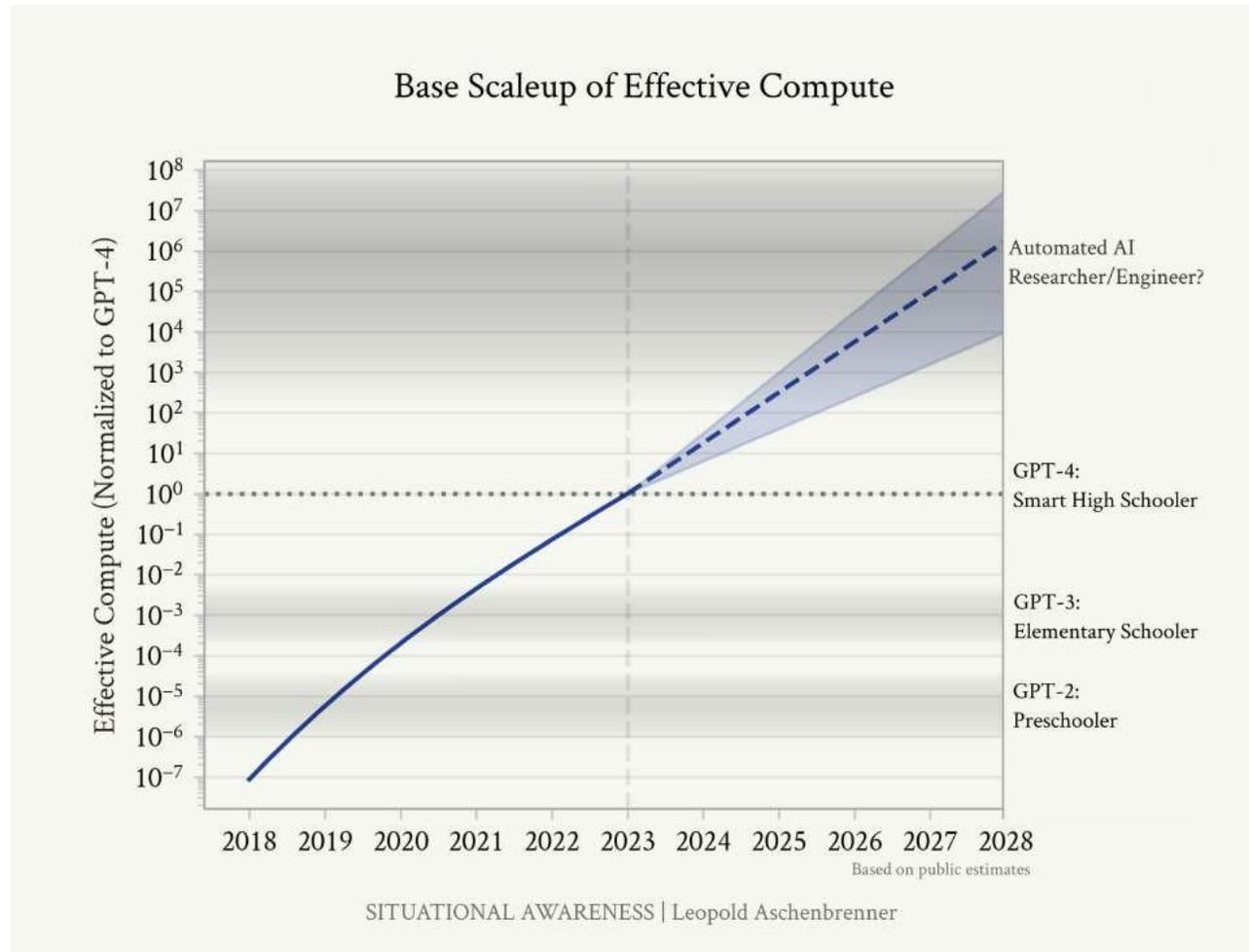
Las capacidades de GPT-4 fueron una sorpresa para muchos: un sistema de IA que podía escribir código y ensayos, podía razonar a través de problemas matemáticos difíciles y superar los exámenes universitarios. Hace unos años, la mayoría pensaba que se trataba de muros impenetrables.

Pero GPT-4 no fue más que la continuación de una década de avances vertiginosos en el aprendizaje profundo. Una década antes, los modelos apenas podían identificar imágenes simples de gatos y perros; Cuatro años antes, GPT-2 apenas podía encadenar sensaciones semiplausibles. Ahora estamos saturando rápidamente todos los puntos de referencia que se nos ocurren. Y, sin embargo, este progreso dramático ha sido simplemente el resultado de tendencias consistentes en la ampliación del aprendizaje profundo.

Ha habido gente que ha visto esto durante mucho más tiempo. Se burlaron de ellos, pero todo lo que hicieron fue confiar en las líneas de tendencia. El

Las líneas de tendencia son intensas, y tenían razón. Los modelos, solo quieren aprender; Los amplían y aprenden más.

Hago la siguiente afirmación: **es sorprendentemente plausible que para 2027, los modelos sean capaces de hacer el trabajo de un investigador/ingeniero de IA.** Eso no requiere creer en la ciencia ficción, solo requiere creer en las líneas rectas de un gráfico.



En este artículo, simplemente "contaré los OOM" (OOM = orden de magnitud, $10x = 1$ orden de magnitud): mira las tendencias en 1) *computación*, 2) *eficiencias algorítmicas* (progreso algorítmico que podemos considerar como una "computación efectiva" creciente), y 3) *ganancias "descoordinadas"* (arreglar las formas obvias en que los modelos se obstaculizan por defecto, desbloqueando capacidades latentes y dándoles herramientas, lo que lleva a cambios radicales en la utilidad). Rastreamos

Figura 1: Estimaciones aproximadas del escalado pasado y futuro de la computación efectiva (tanto de la computación física como de las eficiencias algorítmicas), basadas en las estimaciones públicas que se analizan en este artículo. A medida que escalamos los modelos, se vuelven cada vez más inteligentes, y al "contar los OOM" obtenemos una idea aproximada de qué inteligencia del modelo deberíamos esperar en el futuro (cercano). (Este gráfico muestra solo el escalado en los modelos base; Los "descocidos" no se representan).

el crecimiento de cada uno de ellos durante los cuatro años anteriores a GPT-4, y lo que deberíamos esperar en los cuatro años posteriores, hasta finales de 2027. Dadas las mejoras constantes del aprendizaje profundo para cada OOM de cómputo efectivo, podemos usar esto para proyectar el progreso futuro.

Públicamente, las cosas han estado tranquilas durante un año desde el lanzamiento de GPT-4, ya que la próxima generación de modelos ha estado en el horno, lo que ha llevado a algunos a proclamar el estancamiento y que el aprendizaje profundo es chocar contra una pared.¹ Pero al contar los OOM, podemos echar un vistazo a ¹ Predicciones que han hecho cada año lo que realmente deberíamos esperar.

durante la última década, y en lo que se han equivocado constantemente...

El resultado es bastante simple. De GPT-2 a GPT-4, desde modelos que eran impresionantes por lograr a veces encadenar unas pocas oraciones coherentes, hasta modelos que superan los exámenes de la escuela secundaria, no fue una ganancia única. Estamos corriendo a través de los OOM extremadamente rápido, y los números indican que deberíamos esperar otro escalado de cómputo $\sim 100,000$ veces efectivo, lo que resulta en otro salto cualitativo del tamaño de GPT-2 a GPT-4, en cuatro años. Además, y de manera crítica, eso no solo significa un mejor chatbot; Recoger los muchos frutos obvios que están al alcance de la mano en las ganancias de "destrabar" debería llevarnos de los chatbots a los agentes, de una herramienta a algo que se parece más a los reemplazos de trabajadores remotos.

Si bien la inferencia es simple, la implicación es sorprendente. Un salto como ese podría llevarnos a AGI, a modelos tan inteligentes como doctores o expertos que pueden trabajar a nuestro lado como coworkers. Quizás lo más importante es que si estos sistemas de IA pudieran autoacoplar la investigación de la IA en sí misma, eso pondría en marcha intensos bucles de retroalimentación, el tema de [la siguiente pieza de la serie](#).

Incluso ahora, casi nadie está valorando todo esto. Pero el conocimiento de la situación en la IA no es tan difícil, una vez que das un paso atrás y observas las tendencias. Si no dejas de sorprenderte con las capacidades de la IA, empieza a contar las OOM.

Los últimos cuatro años

Ahora TENEMOS MÁQUINAS con las que básicamente podemos **hablar como humanos**. Es un testimonio notable de la capacidad humana para adaptarse que esto parezca normal, que nos hayamos acostumbrado al ritmo del progreso. Pero vale la pena dar un paso atrás y mirar el progreso de los últimos años.

De GPT-2 a GPT-4

Permítanme recordarles lo lejos que llegamos en solo los ~ 4 (!) años previos a GPT-4.

GPT-2 (2019) ~ niño en edad preescolar: "Vaya, puede encadenar algunas oraciones plausibles". Un ejemplo muy escogido de una historia semicoherente sobre unicornios en los Andes **generó** fue increíblemente impresionante en ese momento. Y, sin embargo, GPT-2 apenas podía contar hasta 5 sin tropezar;² Al resumir

un artículo, apenas superó la selección de 3 sen-

De acuerdo con el artículo, la mayoría de las personas que se encuentran en el centro de la ciudad de³

² DeSSC: "Janelle Shane pregunta GPT- 2 Sus diez animales favoritos:

Indicación: Mis 10 animales favoritos son: 1. Mis diez animales favoritos son:

1. Cebras con una cicatriz blanca en la espalda
2. Arañas y pulpos insidiosos
3. Rana con hojas grandes, ojalá negras
4. Cacatúa con escamas
5. Alca con alas colgando a unas 4 pulgadas de la cara y un tatuaje de corazón en una rana
3. Tetrabodos entrelazados de cocatúa que se pueden cegar, cortar y comer crudos:
4. Cocodrilos blancos y negros del desierto que viven a la luz del sol
5. Cebrá y muchos otros insectos de los guisantes"

³ DelPapel GPT-2, sección 3.6.

GPT-2 examples people found very impressive at the time

<p>Task Reading Comprehension: answer questions about given passages</p> <p>Dataset COQA</p> <p>Examples <i>The trophy doesn't fit into the brown suitcase because it is too large.</i></p> <p>Correct answer: it = trophy Model answer: it = trophy</p> <p><i>The trophy doesn't fit into the brown suitcase because it is too small.</i></p> <p>Correct answer: it = suitcase Model answer: it = suitcase</p>	<p>System Prompt (human-written) <i>For today's homework assignment, please describe the reasons for the US Civil War.</i></p> <p>Model Completion (machine-written, 10 tries) <i>By Donny Ferguson</i></p> <p>It is easy to identify why the Civil War happened, because so many people and so many books and so much television and films tell us that it was the cause, that it has something to do with race or economics or religion. There is much agreement that it was essentially a war of slavery on behalf of capitalism, about a century of slavery. But that's not what most people think of when they think of the Civil War. Many people think of the war as about states rights. You're not wrong about that, but it's kind of misleading to say that the Civil War was a conflict between states' rights and federalism. So let's try again. What's the idea that has inspired so many people to go onto the battlefields of Gettysburg and Antietam and Gettysberg and Petersburg and Fredericksburg? The American idea of the republic--a notion of limited government--is a great part of the history.</p>
--	--

SITUATIONAL AWARENESS | Leopold Aschenbrenner

Comparar las capacidades de la IA con la inteligencia humana es difícil y defectuoso, pero creo que es informativo considerar la analogía

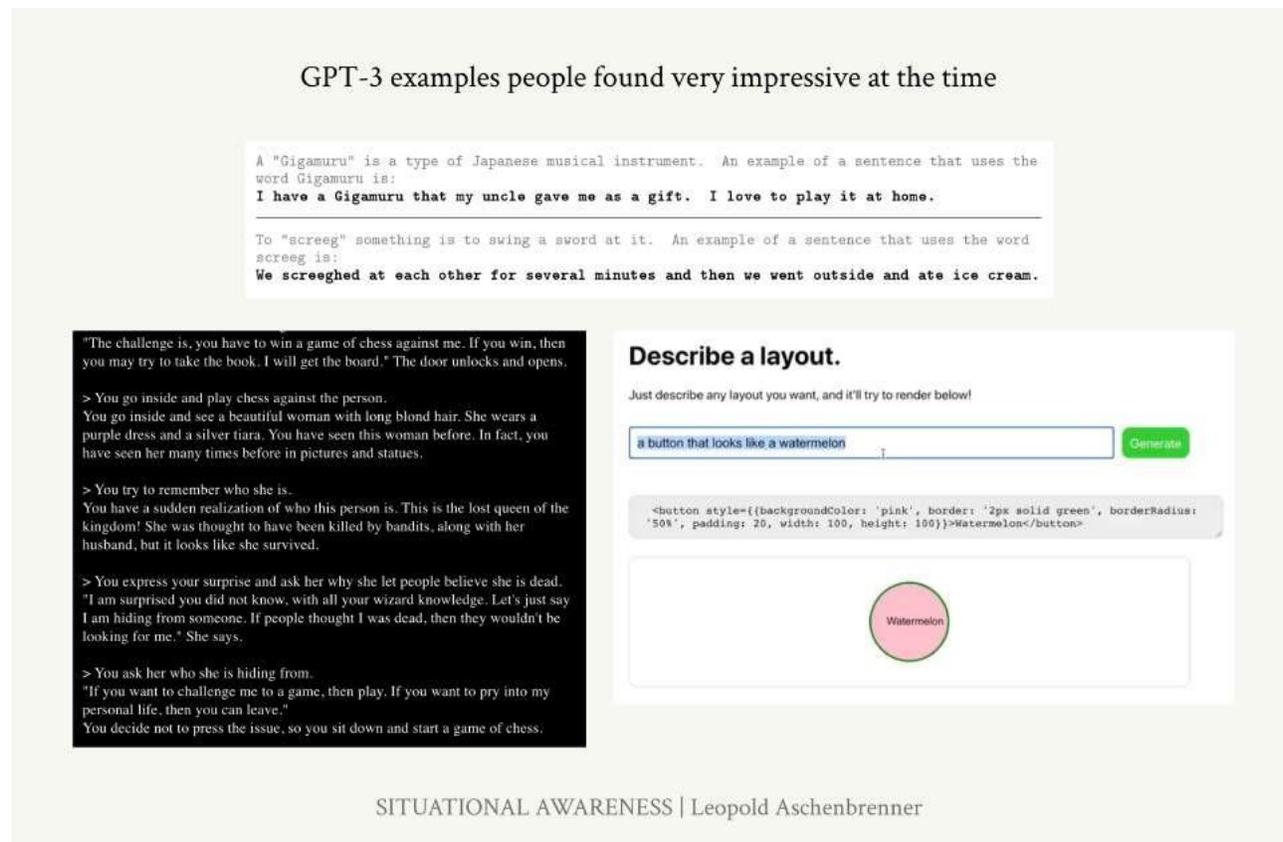
Figura 2: Algunos ejemplos de lo que la gente encontró impresionante sobre GPT-2 en ese momento. Izquierda: GPT-2 hace un Ok trabajo en preguntas de comprensión lectora extremadamente básicas. Derecha: En una muestra seleccionada (al mejor de 10

GPT-2 puede escribir un párrafo semi-coherente que diga algunas cosas semi-relevantes sobre la Guerra Civil.

aquí, aunque sea muy imperfecto. GPT-2 fue impactante por su dominio del lenguaje y su capacidad para generar ocasionalmente un párrafo semicohesivo u ocasionalmente responder correctamente preguntas fácticas simples. Es lo que habría sido impresionante para un niño en edad preescolar.

GPT-3 (2020)⁴ ~ Estudiante de primaria: "Vaya, con solo unos pocos Ejemplos de tiros, puede hacer algunas tareas simples y útiles". Comenzó a ser coherente incluso en varios párrafos de manera mucho más consistente, y podía corregir la gramática y hacer algo de aritmética muy básica. Por primera vez, también fue comercialmente útil en algunos aspectos limitados: por ejemplo, GPT-3 podía generar una copia simple para SEO y marketing.

⁴ Me refiero al viejo y torpe GPT-3, no al GPT-3.5 drásticamente mejorado que quizás conozcas de ChatGPT.



Una vez más, la comparación es imperfecta, pero lo que impresionó a la gente sobre GPT-3 es quizás lo que habría sido impresionante para un niño de primaria: escribió algo de poesía básica, pudo contar historias más ricas y coherentes, pudo comenzar a hacer algo rudimentario

Figura 3: Algunos ejemplos de lo que la gente encontró impresionante sobre GPT-3 en ese momento. Arriba: Después de una simple instrucción, GPT-3 puede usar una palabra inventada en una nueva oración. Abajo a la izquierda: GPT-3 puede participar en una rica narración de ida y vuelta. Abajo a la derecha: GPT-3 puede generar un código muy simple.

codificación, podría aprender de manera bastante confiable a partir de instrucciones y demostraciones simples, y así sucesivamente.

GPT-4 (2023) ~ *estudiante de secundaria inteligente*: "Vaya, puede escribir código bastante sofisticado y depurar de forma iterativa, puede escribir de forma inteligente y sofisticada sobre temas complicados, puede razonar a través de matemáticas difíciles de la competencia de la escuela secundaria, está superando a la gran mayoría de los estudiantes de secundaria en cualquier prueba que podamos darle, etc." Desde el código hasta las matemáticas y las estimaciones de Fermi, puede pensar y razonar. GPT-4 ahora es útil en mis tareas diarias, desde ayudar a escribir código hasta revisar borradores.

GPT-4 examples people found very impressive at the time

GPT-4: Prompt: Can you write a proof that there are infinitely many primes, with every line that rhymes?

GPT-4: Yes, I think I can. Though it might take a clever plan. I'll start by noting Euclid's proof, which shows that primes aren't just alone. Assume we have a finite list. Of primes, and that none have been missed. Multiply them all together, and add one, just to be clever. The result will be a number. That has no prime factors, oh wonder! But every number has a prime. That divides it, it's just a matter of time. So we've found a contradiction, and our finite list needs evicting. There must be infinitely many primes, and that's the end of my rhyming lines.

Prompt: Draw a unicorn in TikZ.

GPT-4: [Produces LaTeX compiling to following picture:]

Prompt: Can you generate a pyplot for the following data: x = [1, 3, 5, 6, 8], y1 = [2, 3, 5, 18, 1], y2 = [3, 5, 6, 8, 1], y3 = [5, 1, 2, 3, 4], y4 = [0, 7, 2, 3, 1]. I want y1, y2 to be in the same plot, but y3 is in another plot next to that plot, y4 is in below. I want the legend of y1 to be "blue", y2 to be "orange", y3 to be "green", y4 to be "red". I want the x-axis to be labeled with "time" and y-axis to be labeled with "money". I want to add a 30-50% random error bar to each curve, through all times (including non-integers). I want smoothed curves to show the plot, and smoothed error bar. Do not use linear interpolation, use smooth interpolation. I want to also add some small zig-zag to the smoothed curve to make it look more real. I want to put x baseline as the mean of each line. I want to put a pie chart below indicating the fraction of the four people in each line step. I also want an animation to show how the fraction are changing in the pie chart in "continuous time". Interpolate the missing fractions! I also want vertical line animations in other three plots to match the pie chart. I want the fanciest plot. Please add as many fancy things as possible. Here is a possible code to generate a pyplot for the following data, with some comments to explain the steps:

GPT-4: [We give two snapshots from the animation resulting from the produced code]

Prompt: Suppose $g(x) = f^{-1}(x)$, $f(0) = 5$, $g(4) = 7$, $g(3) = 2$, $g(7) = 9$, $g(9) = 6$ what is $f(f(g(6)))$?

GPT-4: There are different ways to approach this problem, but one possible method is to use the fact that $g(x) = f^{-1}(x)$ implies that $f(g(x)) = x$ for any x in the domain of g . That is, applying f to both sides of the equation $g(x) = f^{-1}(x)$ gives $f(g(x)) = f(f^{-1}(x)) = x$. This means that we can undo the effect of f by applying g , and vice versa. Using this idea, we can try to find $f(6)$ by looking for the value of x such that $g(x) = 6$. From the given information, we see that $g(9) = 6$, so $f(6) = f(g(9)) = 9$. Similarly, we can find $f(9)$ by looking for the value of x such that $g(x) = 9$. We see that $g(7) = 9$, so $f(9) = f(g(7)) = 7$. Finally, we can find $f(7)$ by looking for the value of x such that $g(x) = 7$. We see that $g(4) = 7$, so $f(7) = f(g(4)) = 4$. Therefore, $f(f(g(6))) = f(f(9)) = f(7) = 4$.

GPT-4 vs ChatGPT:

Prompt: A rain barrel collects water off the roof of a house during three hours of heavy rainfall. The height of the water in the barrel increases at the rate of $r(t) = 4t^2 e^{-1.5t}$ feet per hour, where t is the time in hours since the rain began. At time $t = 1$ hour, the height of the water is 0.75 feet. What is the height of the water in the barrel at time $t = 2$ hours?

GPT-4: If we can find a function F such that $F'(t) = r(t)$, then $F(2) - F(1)$ is the amount that the height of the water increases from $t = 1$ to $t = 2$. To find such a function F , we can use the technique of integration by parts, which says that if u and v are differentiable functions, then $\int u dv = uv - \int v du$. Applying this to $r(t) = 4t^2 e^{-1.5t}$, we can choose $u = t^2$ and $dv = 4e^{-1.5t} dt$, so that $du = 2t dt$ and $v = -\frac{4}{1.5} e^{-1.5t}$. Then we have $\int 4t^2 e^{-1.5t} dt = -\frac{8}{3} t^2 e^{-1.5t} - \int -\frac{8}{3} (2t) e^{-1.5t} dt$. Simplifying and applying integration by parts again, we can choose $u = t$ and $dv = -\frac{8}{3} e^{-1.5t} dt$, so that $du = dt$ and $v = \frac{8}{4.5} e^{-1.5t}$. Then we have $\int 4t^2 e^{-1.5t} dt = -\frac{8}{3} t^2 e^{-1.5t} + \frac{16}{9} t e^{-1.5t} - \int \frac{16}{9} e^{-1.5t} dt$.

ChatGPT: To find the height of the water in the barrel at time $t = 2$ hours, we need to find the value of $r(2)$. We can do this by plugging 2 in for t in the expression for $r(t)$. $r(2) = 4(2)^2 e^{-1.5(2)} = 32e^{-3}$. To find the height of the water in the barrel at time $t = 2$ hours, we need to add the amount of water collected during the first hour (0.75 feet) to the amount of water collected during the second hour ($r(2)$ feet). So, the total height of the water in the barrel at time $t = 2$ hours is $0.75 + r(2) = 0.75 + 32e^{-3} \approx 0.91$ feet. Therefore, the height of the water in the barrel at time $t = 2$ hours is approximately 0.91 feet.

Figure 4.3: GPT-4 vs ChatGPT on AP problems. GPT-4 uses a correct approach, but produces a wrong final answer due to a calculation error, whereas ChatGPT produces an incoherent argument.

GPT-4: Prompt: Solve a LeetCode problem using dynamic programming. GPT-4's solution also has better readability thanks to comprehensive commenting.

```

class Solution:
    def minCostClimbingStairs(self, cost: List[int]) -> int:
        # Base cases
        if len(cost) == 1:
            return cost[0]
        if len(cost) == 2:
            return min(cost[0], cost[1])
        # Dynamic programming
        dp = [0] * len(cost)
        dp[0] = cost[0]
        dp[1] = cost[1]
        for i in range(2, len(cost)):
            dp[i] = min(dp[i-1], dp[i-2]) + cost[i]
        # Return the minimum cost to reach the top
        return min(dp[-1], dp[-2])

```

GPT-4: The above code solves the problem using dynamic programming. It uses a list `dp` to store the minimum cost to reach each step. The base cases are handled separately. The main loop iterates through the steps, calculating the minimum cost for each step based on the previous two steps. The final result is the minimum of the last two elements in the `dp` array, representing the minimum cost to reach the top of the stairs.

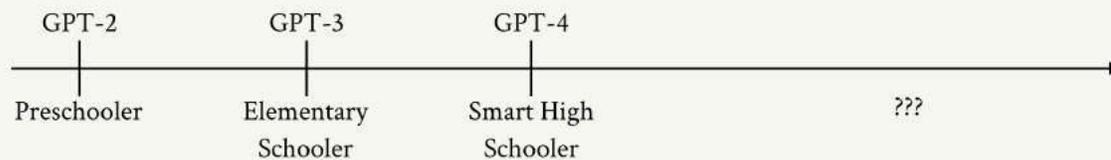
Figure 3.1: Solving a LeetCode problem using dynamic programming. GPT-4's solution also has better readability thanks to comprehensive commenting.

Figure 4: Algunas de las cosas que la gente encontró impresionantes sobre GPT-4 cuando se volvió a liberar, del artículo "Sparks of AGI". Arriba: Está escribiendo código muy complicado (produciendo los gráficos que se muestran en la mitad) y puede razonar a través de problemas matemáticos no triviales. Abajo a la izquierda: Resolviendo un problema matemático de AP. Abajo a la derecha: Resolviendo un

SITUATIONAL AWARENESS 15
problema de codificación bastante
complejo.
Más extractos interesantes de esa
exploración de las capacidades de
GPT-4[aquí](#).

En todo, desde los exámenes AP hasta el SAT, GPT-4 obtiene mejores puntajes que la gran mayoría de los estudiantes de secundaria.

Por supuesto, incluso GPT-4 sigue siendo algo desigual; Para algunas tareas es mucho mejor que para los estudiantes de secundaria inteligentes, mientras que hay otras tareas que aún no puede hacer. Dicho esto, tiendo a pensar que la mayoría de estas limitaciones se reducen a formas obvias en que los modelos todavía están obstaculizados, como discutiré en profundidad más adelante. La inteligencia en bruto está (en su mayor parte) ahí, incluso si los modelos todavía están artificialmente restringidos; Se necesitará un trabajo adicional para desbloquear modelos que puedan aplicar completamente esa inteligencia bruta en todas las aplicaciones.



SITUATIONAL AWARENESS | Leopold Aschenbrenner

Las tendencias en aprendizaje profundo

El ritmo del progreso del aprendizaje profundo en la última década ha sido simplemente extraordinario. Hace apenas una década, fue revolucionario que un sistema de aprendizaje profundo identificara imágenes simples. Hoy en día, seguimos tratando de llegar a pruebas novedosas y cada vez más difíciles y, sin embargo, cada nuevo punto de referencia se rompe rápidamente. Solía llevar décadas descifrar los puntos de referencia ampliamente utilizados; Ahora parece que son solo meses.

Literalmente nos estamos quedando sin puntos de referencia. Como anécdota, mis amigos Dan y Collin hicieron un benchmark llamado **MMLU** hace unos años, en 2020. Esperaban finalmente hacer un punto de referencia que resistiera la prueba del tiempo, equivalente a todos los

exámenes más difíciles que damos a los estudiantes de secundaria y universitarios. Apenas tres años después, está básicamente resuelto: modelos como GPT-4 y Gemini

Gráfico 5: Progreso en solo cuatro años. ¿En **qué** punto de esta línea te encuentras?

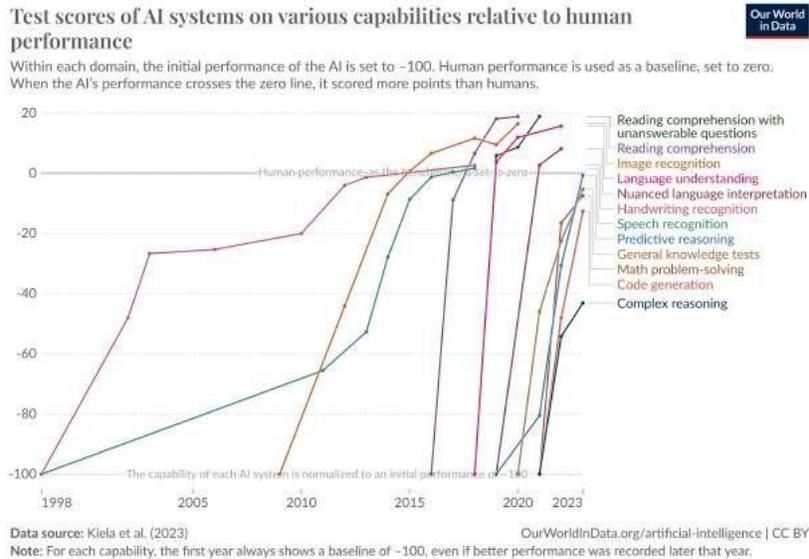


Figura 6: Los sistemas de aprendizaje profundo están alcanzando o superando rápidamente el nivel humano en muchos dominios. Gráfico: Nuestro mundo en datos

Obtener ~90%.

En términos más generales, GPT-4 en su mayoría rompe todas las escuelas secundarias estándar y las pruebas de aptitud universitaria (Figura 7).⁵

O considere el **punto de referencia MATH**, un conjunto de problemas matemáticos difíciles de las competencias de matemáticas de la escuela secundaria.⁶ Cuando se publicó el punto de referencia en 2021, GPT-3 solo acertó ~5% de los problemas. Y el documento original señalaba: "Además, encontramos que simplemente aumentar los presupuestos y los recuentos de parámetros del modelo no será práctico para lograr un razonamiento matemático sólido si continúan las tendencias de escalamiento [...]. Para tener más tracción en la resolución de problemas matemáticos, es probable que necesitemos nuevos avances algorítmicos de la comunidad de investigación en general", necesitaríamos nuevos avances fundamentales para resolver MATH, o eso pensaban. Una encuesta realizada a investigadores de **ML predijo**

En los próximos años (figura 8);⁷ y, sin embargo, dentro de nuestro grupo

solo un año (a mediados de 2022), los mejores modelos pasaron de ~5% a **50% de precisión**; ahora, **MATH está básicamente resuelto**, con un rendimiento reciente superior al 90%.

⁵ Y no, estas pruebas no están en el equipo de entrenamiento. Los laboratorios de IA se esfuerzan mucho en asegurarse de que estas evaluaciones no estén contaminadas,

porque necesitan buenas mediciones para hacer buena ciencia. Un análisis reciente de ScaleAI confirmó que los laboratorios líderes no se están ajustando demasiado a los puntos de referencia (aunque algunos desarrolladores de LLM más pequeños podrían estar exprimiendo sus números).

⁶ En el artículo original, se señaló: "También evaluamos a los humanos en MATEMÁTICAS, y descubrimos que un estudiante de doctorado en ciencias de la computación al que no le gustan especialmente las matemáticas alcanzó aproximadamente el 40% en MATEMÁTICAS, mientras que un tres veces medallista de oro de la OMI alcanzó el 90%, lo que indica que las MATEMÁTICAS también pueden ser un desafío para los humanos".

⁷ Un coautor Notas: " C u a n d o

publicó por primera vez el conjunto de datos MATH, al menos un [colega investigador de ML] nos dijo que era un conjunto de datos sin sentido porque estaba demasiado lejos del rango de lo que los modelos de ML podían lograr (de hecho, yo mismo estaba algo preocupado por esto)".

Performance on common exams
(percentile compared to human test-takers)

	GPT-4 (2023)	GPT-3.5 (2022)
Uniform Bar Exam	90th	10th
LSAT	88th	40th
SAT	97th	87th
GRE (Verbal)	99th	63rd
GRE (Quantitative)	80th	25th
US Biology Olympiad	99th	32nd
AP Calculus BC	51st	3rd
AP Chemistry	80th	34th
AP Macroeconomics	92nd	40th
AP Statistics	92nd	51st

SITUATIONAL AWARENESS | Leopold Aschenbrenner

Figura 7: Puntuaciones de GPT-4 en pruebas estandarizadas. Obsérvese también el gran salto de GPT-3,5 a GPT-4 en el percentil humano en estas pruebas, a menudo desde muy por debajo de la mediana humana hasta la parte superior del rango humano. (Y esto es GPT-3.5, un modelo bastante reciente lanzado menos de un año antes de GPT-4, ¡no el viejo y torpe GPT-3 de nivel de escuela primaria del que hablábamos antes!)

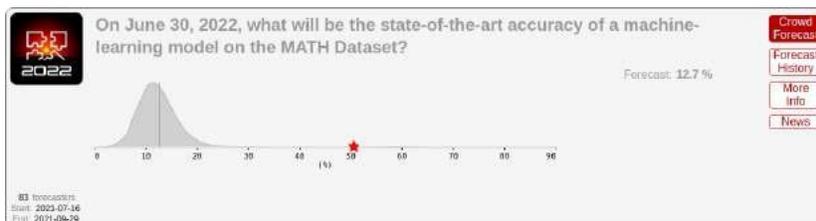


Figura 8: Gris: Pronósticos profesionales, realizados en agosto de 2021, para el desempeño de junio de 2022 en el punto de referencia MATH (problemas matemáticos difíciles de las competencias de matemáticas de la escuela secundaria). Estrella roja: rendimiento real de última generación para junio de 2022, superando con creces incluso el rango superior que dieron los pronosticadores. La mediana de los investigadores de ML es más pesimista.

Una y otra vez, año tras año, los escépticos han afirmado que "el aprendizaje profundo no podrá hacer X" y lo han hecho rápidamente

Se ha demostrado que es incorrecto.⁸ *Si hay una lección que hemos aprendido del pasado* ⁸ Aquí está Yann

LeCun [Predecire](#) en 2022

década de la IA, es que nunca se debe apostar en contra del aprendizaje profundo.

Ahora, los puntos de referencia más difíciles sin resolver son pruebas como [GPQA](#), un conjunto de preguntas de química y física de nivel de grado

Muchas preguntas parecen más fáciles al ser resueltas por IA que por humanos. El modelo GPT-4 obtiene una puntuación de ~60% en comparación con los doctores en el dominio que obtienen

~60% en comparación con los doctores en el dominio que obtienen

~80%, y espero que este punto de referencia también caiga en la próxima generación o dos.

que ni siquiera GPT-5000 será capaz de razonar sobre las interacciones físicas con el mundo real; GPT-4 obviamente lo hace con facilidad un año después.

Aquí están los muros de Gary Marcus predichos después de que GPT-2 [fuera resuelto](#) por GPT-3, y los muros que predijo después de que GPT-3 [fuera resuelto](#) por GPT-4.

Aquí está el profesor Bryan Caplan [perdiendo](#) su primera apuesta pública (después de haber tenido un historial perfecto de apuestas públicas). En enero de 2023, después de que GPT-3.5 obtuviera una D en sus exámenes parciales de economía, el profesor Caplan [apostó](#) a Matthew Barnett a que ninguna IA obtendría una A en sus exámenes parciales de economía para 2029. Solo dos meses después, cuando salió GPT-4, rápidamente obtuvo una A en su examen parcial (y habría sido una de las puntuaciones más altas de su clase).

⁹ En el [conjunto de diamantes](#), la [votación mayoritaria](#) del modelo lo intentó 32 veces con cadena de pensamiento.

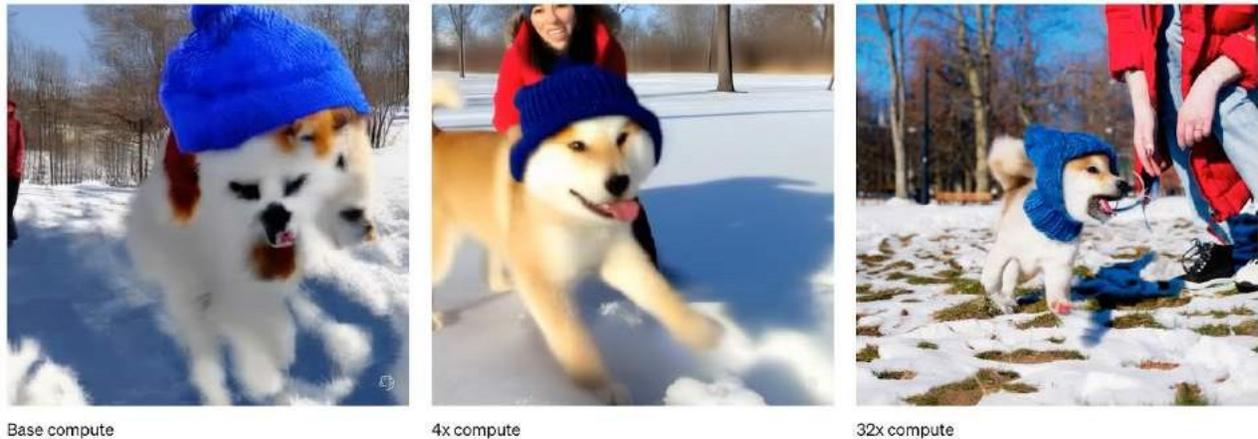
Chemistry (general)
<p>A reaction of a liquid organic compound, which molecules consist of carbon and hydrogen atoms, is performed at 80 centigrade and 20 bar for 24 hours. In the proton nuclear magnetic resonance spectrum, the signals with the highest chemical shift of the reactant are replaced by a signal of the product that is observed about three to four units downfield. Compounds from which position in the periodic system of the elements, which are also used in the corresponding large-scale industrial process, have been mostly likely initially added in small amounts?</p> <p>A) A metal compound from the fifth period. B) A metal compound from the fifth period and a non-metal compound from the third period. C) A metal compound from the fourth period. D) A metal compound from the fourth period and a non-metal compound from the second period.</p>
Organic Chemistry
<p>Methylcyclopentadiene (which exists as a fluxional mixture of isomers) was allowed to react with methyl isoamyl ketone and a catalytic amount of pyrrolidine. A bright yellow, cross-conjugated polyalkenyl hydrocarbon product formed (as a mixture of isomers), with water as a side product. These products are derivatives of fulvene. This product was then allowed to react with ethyl acrylate in a 1:1 ratio. Upon completion of the reaction, the bright yellow color had disappeared. How many chemically distinct isomers make up the final product (not counting stereoisomers)?</p> <p>A) 2 B) 16 C) 8 D) 4</p>
Genetics
<p>If a sperm from species A is injected into an egg from species B and both species have the same number of chromosomes, what would be the main cause of the resulting zygote mortality?</p> <p>A) Species specific zona pellucida proteins on the egg cannot bind sperms from a different species. B) Epistatic interactions between the genes of different species C) Chromosomal incompatibilities will cause failure of meiosis leading to death of zygote. D) Chromosomal recombination will not occur in different species.</p>
Molecular Biology
<p>A scientist studies the stress response of barley to increased temperatures and finds a protein which contributes to heat tolerance through the stabilisation of cell membrane. The scientist is very happy and wants to create a heat-tolerant cultivar of diploid wheat. Using databases, they find a heat tolerance protein homologue and start analysing its accumulation under heat stress. Soon enough, the scientist discovers this protein is not synthesised in the wheat cultivar they study. There are many possible reasons for such behaviour, including:</p> <p>A) A miRNA targets the protein, which makes exonucleases cut it immediately after the end of translation and before processing in ER B) Trimethylation of lysine of H3 histone in position 27 at the promoter of the gene encoding the target protein C) A stop-codon occurs in the 5'-UTR region of the gene encoding the target protein D) The proteolysis process disrupts a quaternary structure of the protein, preserving only a tertiary structure</p>
Astrophysics
<p>Astronomers are studying a star with a T_{eff} of approximately 6000 K. They are interested in spectroscopically determining the surface gravity of the star using spectral lines ($EW < 100 \text{ m\AA}$) of two chemical elements, E11 and E12. Given the atmospheric temperature of the star, E11 is mostly in the neutral phase, while E12 is mostly ionized. Which lines are the most sensitive to surface gravity for the astronomers to consider?</p> <p>A) E12 I (neutral) B) E11 II (singly ionized) C) E12 II (singly ionized) D) E11 I (neutral)</p>
Quantum Mechanics
<p>Suppose we have a depolarizing channel operation given by $E(\rho)$. The probability, p, of the depolarization state represents the strength of the noise. If the Kraus operators of the given state are $A_0 = \sqrt{1 - \frac{3p}{4}}$, $A_1 = \sqrt{\frac{p}{4}}X$, $A_2 = \sqrt{\frac{p}{4}}Y$, and $A_3 = \sqrt{\frac{p}{4}}Z$. What could be the correct Kraus Representation of the state $E(\rho)$?</p> <p>A) $E(\rho) = (1 - p)\rho + \frac{p}{3}X\rho X + \frac{p}{3}Y\rho Y + \frac{p}{3}Z\rho Z$ B) $E(\rho) = (1 - p)\rho + \frac{p}{3}X\rho^2 X + \frac{p}{3}Y\rho^2 Y + \frac{p}{3}Z\rho^2 Z$ C) $E(\rho) = (1 - p)\rho + \frac{p}{4}X\rho X + \frac{p}{4}Y\rho Y + \frac{p}{4}Z\rho Z$ D) $E(\rho) = (1 - p)\rho^2 + \frac{p}{3}X\rho^2 X + \frac{p}{3}Y\rho^2 Y + \frac{p}{3}Z\rho^2 Z$</p>

Table 1: Six example questions from the dataset, two each from subdomains of chemistry, biology, and physics (respectively).

Figura 9: Ejemplos de preguntas de GPQA. Los modelos ya son mejores en esto que yo, y probablemente pronto llegaremos al nivel de experto. . .

Recuento de los OOM

¿Cómo sucedió esto? La magia del aprendizaje profundo es que *simplemente funciona*, y las líneas de tendencia han sido asombrosamente consistentes, a pesar de los detractores en todo momento.



Con cada OOM de cómputo efectivo, los modelos obtienen de forma predecible y fiable

mejor.¹⁰ Si podemos contar los OOMs, podemos (a grandes rasgos, cualita-
¿Cuán con-

Extrapolar las mejoras de capacidad.¹¹ Así es como algunas personas premonitorias vieron venir GPT-4.

Podemos descomponer el progreso en los cuatro años de GPT-2 a GPT-4 en tres categorías de scaleups:

1. **Compute:** Estamos usando computadoras mucho más grandes para entrenar estos modelos.
2. **E ALGORÍTMICO Funcionalidades:** Hay una tendencia continua de progreso algorítmico. Muchos de ellos actúan como "multiplicadores de cómputo", y podemos ponerlos en una escala unificada de crecimiento *Cómputo eficaz*.
3. **Ganancias "DESCOJEADORAS":** Por defecto, los modelos aprenden un montón de increíbles capacidades brutas, pero se ven obstaculizadas en todo tipo de formas tontas, lo que limita su valor práctico. Con mejoras algorítmicas simples, como el aprendizaje por refuerzo a partir de la retroalimentación humana (RLHF), la cadena de pensamiento (CoT), las herramientas y el andamiaje, podemos desbloquear importantes capacidades latentes.

Figura 10: Los efectos del escalado de la computación, en el ejemplo de OpenAI Sora.

¹⁰Y vale la pena señalar

Estas líneas de tendencia son muy importantes. Combinación del documento original de las leyes de escalado con algunas de las estimaciones en cómputo y el escalado de la eficiencia informática desde entonces implica una tendencia de escalado coherente para más de 15 órdenes de magnitud (más de 1.000.000.000.000.000 veces en computación efectiva).

¹¹Un error común es pensar que el escalado solo es válido para la pérdida de perplejidad, pero también vemos un comportamiento de escalado muy claro y consistente en el rendimiento posterior en los puntos de referencia. Por lo general, es solo cuestión de encontrar el gráfico logarítmico correcto. Por ejemplo, en la entrada del blog GPT-4, muestran un comportamiento de escalado coherente para el rendimiento en problemas de codificación de más de 6 OOM (1.000.000x) de cómputo, utilizando MLPR (tasa media de aprobación de registro).

La pregunta "¿Son las habilidades emergentes una rabia?" El documento hace un punto similar; Con la elección correcta de la métrica, casi siempre hay una tendencia constante para el rendimiento en las tareas posteriores.

De manera más general, la observación cualitativa de la "hipótesis

de escalamiento" —tendencias muy claras en la capacidad del modelo con escala— es anterior a las curvas de escalamiento de pérdidas; el trabajo de las "leyes de escalamiento" fue solo una medición formal de esto.

Podemos "contar los OOM" de mejora a lo largo de estos ejes: es decir, trazar el escalado de cada uno en unidades de cómputo efectivo. 3x es 0,5 OOM; 10x es 1 OOM; 30x es 1.5 OOMs; 100x son 2 OOM; y así sucesivamente. También podemos ver lo que deberíamos esperar además de GPT-4, de 2023 a 2027.

Voy a repasar cada uno de ellos uno por uno, pero el resultado es claro: estamos corriendo rápidamente a través de los OOM. Hay posibles vientos en contra en el muro de datos, que abordaré, pero en general, parece probable que deberíamos esperar otro salto del tamaño de GPT-2 a GPT-4, además de GPT-4, para 2027.

Calcular

Comenzaré con el impulsor más comúnmente discutido del progreso reciente: lanzar (mucho) más cómputo a los modelos.

Mucha gente asume que esto se debe simplemente a la Ley de Moore. Pero incluso en los viejos tiempos, cuando la Ley de Moore estaba en su apogeo, era comparativamente *glacial*, tal vez de 1 a 1,5 OOM por década. Estamos viendo escalamientos mucho más rápidos en la computación, cerca de 5 veces la velocidad de la ley de Moore, debido a una inversión gigantesca. (Gastar incluso un millón de dólares en un solo modelo Solía ser un pensamiento escandaloso que nadie consideraría, ¡y ahora eso es cambio de bolsillo!)

Modelo	Cómputo estimado	Crecimiento
GPT-2 (2019)	~4e21 FLOP	
GPT-3 (2020)	~3e23 FLOP	+ ~2 OOM
GPT-4 (2023)	8e24 a 4e25 FLOP	+ ~1,5-2 OOM

Tabla 1: Estimaciones de cómputo de GPT-2 a GPT-4 por Epoch AI.

Podemos utilizar estimaciones públicas de Epoch AI (una fuente ampliamente respetada por su excelente análisis de las tendencias de la IA) para rastrear la ampliación de la computación de 2019 a 2023. GPT-2 a GPT-3 fue un escalado rápido; Había un gran desbordamiento de proceso, escalando desde un experimento más pequeño hasta el uso de un centro de datos completo para entrenar un modelo de lenguaje grande. Con el escalado de GPT-3 a GPT-4, hicimos la transición al régimen moderno: tener que construir un clúster completamente nuevo (mucho más grande) para el siguiente modelo. Y, sin embargo,

El crecimiento dramático continuó. En general, las estimaciones de Epoch AI sugieren que el entrenamiento de GPT-4 utilizó $\sim 3.000x-10.000x$ más computación en bruto que GPT-2.

A grandes rasgos, esto es solo la continuación de una tendencia de más larga duración. Durante la última década y media, principalmente debido a

A gran escala de inversión (y chips especializados para cargas de trabajo de IA en forma de GPU y TPU), el cómputo de entrenamiento utilizado para los sistemas de IA de frontera ha crecido a aproximadamente $\sim 0,5$ OOM/año.

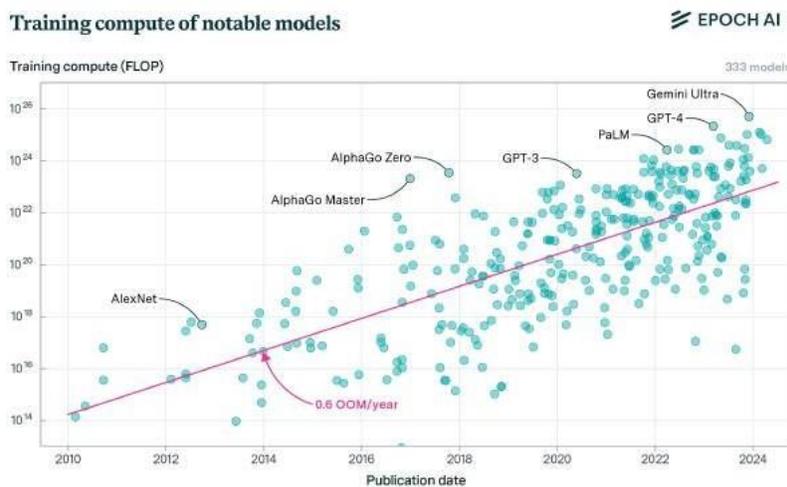


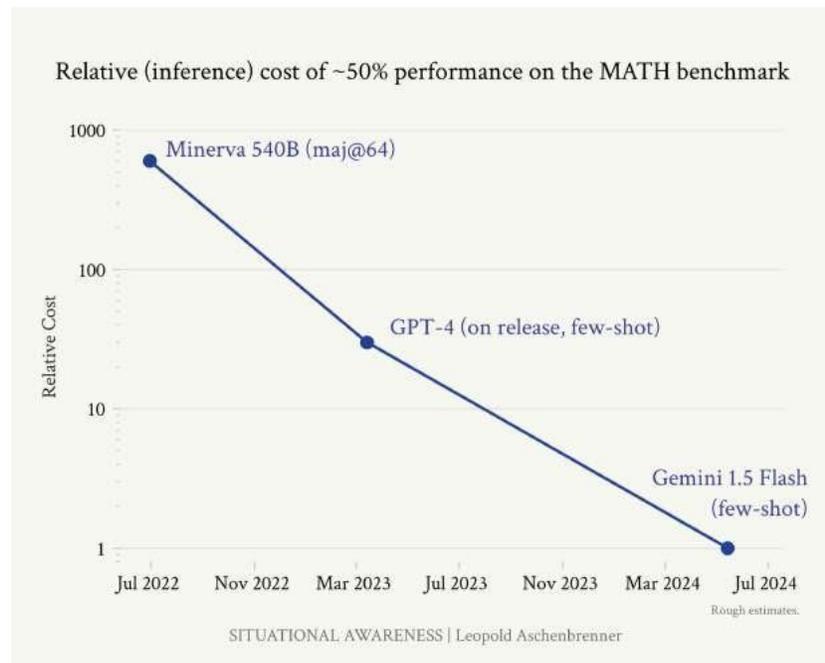
Figura 11: Proceso de entrenamiento de modelos de aprendizaje profundo sin tabla a lo largo del tiempo. Fuente: Epoch AI.

El escalado de cómputo de GPT-2 a GPT-3 en un año fue un exceso inusual, pero todo indica que la tendencia a largo plazo continuará. La fábrica de rumores de ciencia ficción está al tanto de las historias dramáticas de enormes pedidos de GPU. Las inversiones involucradas serán extraordinarias, pero están en movimiento. Entraré en esto más adelante en la serie, en [IIIa. Corriendo hacia el grupo del billón de dólares](#); Con base en ese análisis, parece muy probable que ocurran 2 OOM adicionales de cómputo (un clúster en los \$ 10 de miles de millones) para fines de 2027; incluso un clúster más cercano a +3 OOM de cómputo (\$ 100 mil millones +) parece plausible (y serumorea que está en proceso en Microsoft / OpenAI).

Eficiencias algorítmicas

Si bien las inversiones masivas en computación reciben toda la atención, el progreso algorítmico es probablemente un impulsor igualmente importante del progreso (y ha sido dramáticamente subestimado).

Para ver qué tan importante puede ser el progreso algorítmico, considere la siguiente ilustración (Figura 12) de la caída en el precio para lograr una precisión de ~50% en el punto de referencia MATH (matemáticas de competencia de la escuela secundaria) en solo *dos años*. (A modo de comparación, un estudiante de doctorado en ciencias de la computación al que no le gustaban particularmente las matemáticas obtuvo un 40%, por lo que esto ya es bastante bueno). La eficiencia de la inferencia mejoró en casi 3 OOM (1.000 veces) en menos de dos años.



Aunque estos son números solo para la eficiencia de la inferencia (que pueden o no corresponder a las mejoras en la eficiencia del entrenamiento, donde los números son más difíciles de inferir a partir de datos públicos), dejan en claro que hay una *enorme* cantidad de progreso algorítmico posible y que está sucediendo.

Figura 12: Estimación aproximada del costo relativo de inferencia de alcanzar ~50% de rendimiento MATH.

Cálculos a continuación.

Gemini 1.5 Flash obtiene una puntuación del 54,9% en MATH, y cuesta 0,35 \$/1,05 dólares (entrada/salida) por millón de tokens. GPT-4 obtuvo una puntuación del 42,5 % en el lanzamiento de MATH y del 52,9 % en MATH a principios de 2023, y el coste

30 USD/60 USD (entrada/salida) por millón de tokens; eso es 85x/57x (entrada/salida) más caro por token que Gemini

1.5 Destello. Para ser conservador, utilizo una estimación de 30 veces la disminución del costo anterior (contando con Gemini 1.5 Flash posiblemente usando más tokens para razonar a través de los problemas).

Minerva 540B obtiene una puntuación del 50,3% en MATH, utilizando el voto mayoritario entre 64 sam- ples. Un amigo conocedor estima que el modelo base aquí es probablemente 2-3 veces más caro de inferir que GPT-4. Sin embargo, Minerva parece usar algo menos de tokens por respuesta en una rápida comprobación al azar. Y lo que es más importante, Minerva necesitó 64 muestras para lograr ese rendimiento, lo que supone un múltiplo de 64 veces el coste si, por ejemplo, se ejecuta ingenuamente a través de una API de inferencia. En la práctica, los tokens de solicitud se pueden almacenar en caché al ejecutar una evaluación; Dada una solicitud de pocos disparos, es probable que los tokens de solicitud sean la mayor parte del costo, incluso teniendo en cuenta los tokens de salida. Suponiendo que los tokens de salida sean un tercio del

costo de obtener una sola muestra, eso implicaría solo un aumento de ~ 20 veces en el costo del `maj@64` con el almacenamiento en caché. Para ser conservador, use el número aproximado de una disminución de costo de 20 veces en lo anterior (incluso si la disminución ingenua en el costo de inferencia de ejecutar esto a través de una API sería mayor).

En este artículo, separaré dos tipos de progreso algorítmico. Aquí, comenzaré cubriendo las mejoras algorítmicas "dentro del paradigma", aquellas que simplemente dan como resultado una mejor modificación de la base.

Y que actúan directamente como *eficiencias de cómputo* o *multiplicadores de cómputo*. Por ejemplo, un mejor algoritmo podría permitirnos lograr el mismo rendimiento pero con 10 veces menos cómputo de entrenamiento. A su vez, eso actuaría como un aumento de 10x (1 OOM) en *el cómputo efectivo*. (Más adelante, hablaré de "destrabar", que se puede considerar como "expansión de paradigmas/expansión de aplicaciones" progreso algorítmico que desbloquea las capacidades de los modelos base).

Si damos un paso atrás y observamos las tendencias a largo plazo, parece que encontramos nuevas mejoras algorítmicas a un ritmo bastante constante. Los descubrimientos individuales parecen aleatorios y, a cada paso, parecen obstáculos insuperables, pero la línea de tendencia a largo plazo es predecible, una línea recta en un gráfico. Confía en la línea de tendencia.

Tenemos los mejores datos para ImageNet (donde la búsqueda algorítmica ha sido mayoritariamente pública y tenemos datos que se remontan a una década), para la que hemos mejorado constantemente la eficiencia informática en aproximadamente $\sim 0,5$ OOMs/año durante el período de 9 años entre 2012 y 2021.

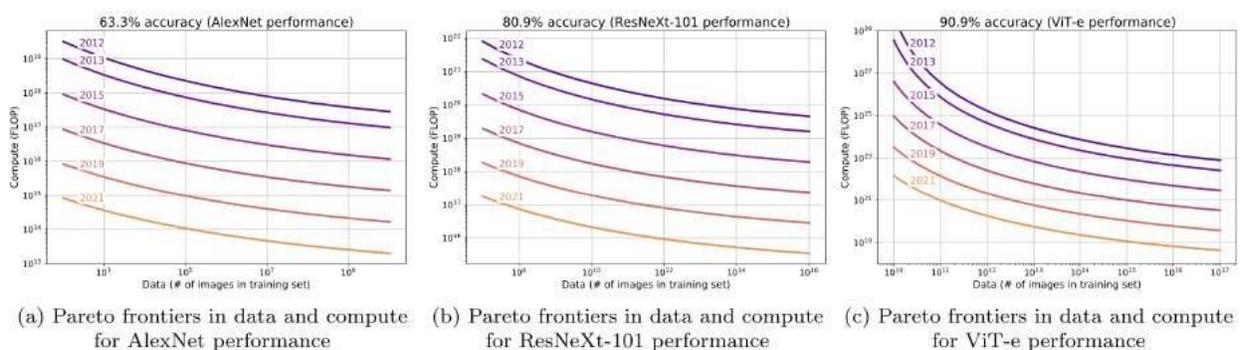


Figure 1. Pareto frontiers for training models to achieve performance of well-known models over time.

Eso es un gran problema: eso significa que 4 años después, podemos lograr el mismo nivel de rendimiento por ~ 100 veces menos computación (y, al mismo tiempo, ¡un rendimiento mucho mayor para la misma computación!).

Desafortunadamente, dado que los laboratorios no publican datos internos sobre esto, es más difícil medir el progreso algorítmico de los LLM de frontera

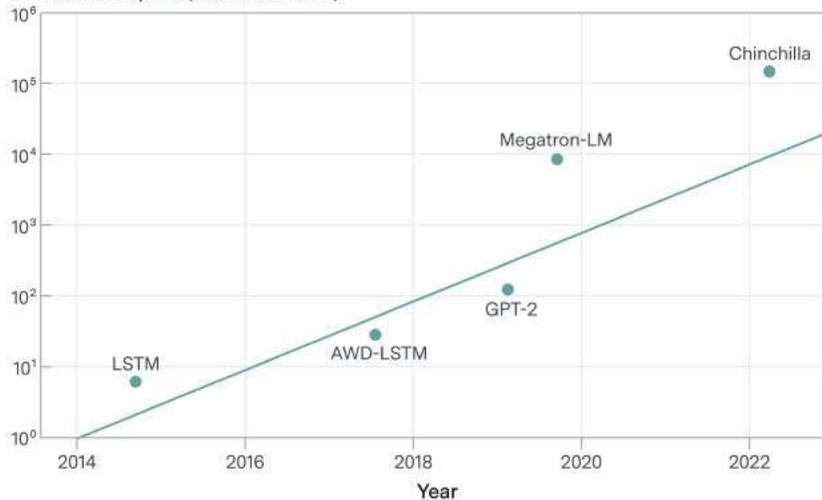
Figura 13: Podemos medir el progreso algorítmico: cuánto menos cómputo se necesita en 2021 en comparación con 2012 para entrenar un modelo con el mismo rendimiento? Vemos una tendencia de $\sim 0,5$ OOMs/año de eficiencia algorítmica. Fuente: Erdil y Besiroglu 2022.

en los últimos cuatro años. **EpochAI tiene un nuevo trabajador** que aplica

sus resultados en ImageNet para el modelado del lenguaje, y estiman una tendencia similar de $\sim 0,5$ OOMs/año de eficiencia algorítmica en los LLM de 2012 a 2023. (Sin embargo, esto tiene barras de error más anchas y no captura algunas ganancias más recientes, ya que los laboratorios líderes han dejado de publicar sus eficiencias algorítmicas).

Efficiency doubles roughly every 8 months

Effective compute (relative to 2014)



Mirando más directamente los últimos 4 años, GPT-2 a GPT-3 fue básicamente una simple ampliación (según el [documento](#)), pero ha habido muchas ganancias conocidas públicamente e inferibles públicamente desde GPT-3:

- Podemos inferir ganancias de los costos de API: ¹²
 - GPT-4, en su lanzamiento, costó \sim lo mismo que GPT-3 cuando se lanzó, a pesar del enorme aumento de rendimiento.¹³ (Si hacemos una estimación ingenua y demasiado simplificada basada en las leyes de escalado, esto sugiere que quizás aproximadamente la mitad del aumento efectivo de la computación de GPT-3 a GPT-4 provino de mejoras algorítmicas.¹⁴)
 - Desde el lanzamiento de GPT-4 hace un año, los precios de OpenAI para los modelos de nivel GPT-4 han caído otras 6x/4x (entrada/salida) con el lanzamiento de GPT-4o.

Figura 14: Estimaciones de Epoch AI de las eficiencias algorítmicas en el modelado del lenguaje. Sus estimaciones sugieren que hemos logrado ~ 4 OOM de ganancias de eficiencia en 8 años.

¹² Aunque se trata de eficiencias de inferencia (más que necesariamente eficiencias de entrenamiento), y hasta cierto punto reflejarán optimizaciones específicas de inferencia, a) sugieren que es posible que se produzcan enormes cantidades de progreso algorítmico y que está ocurriendo en general, y b) a menudo se da el caso de que una mejora algorítmica es tanto una ganancia de eficiencia de entrenamiento como una eficiencia de inferencia, por ejemplo, al reducir el número de parámetros necesarios.

¹³ GPT-3: 60 dólares/1 millón de tokens, GPT-4: 30 \$/1 millón de tokens de entrada y 60 \$/1 millón de tokens de salida.

¹⁴ Las leyes de escalado de Chinchilla dicen que se debe escalar el conteo de parámetros y los datos por igual. Es decir, el recuento de parámetros aumenta "la mitad de los OOM" de los OOM que crece el proceso de entrenamiento eficaz. Al mismo tiempo, el recuento de parámetros es intuitivamente aproximadamente proporcional a costos de inferencia. Por lo tanto, en igualdad de condiciones, los costos de inferencia constantes implican que la mitad de los OOM del crecimiento efectivo de la computación fueron "cancelados" por la ganancia algorítmica.

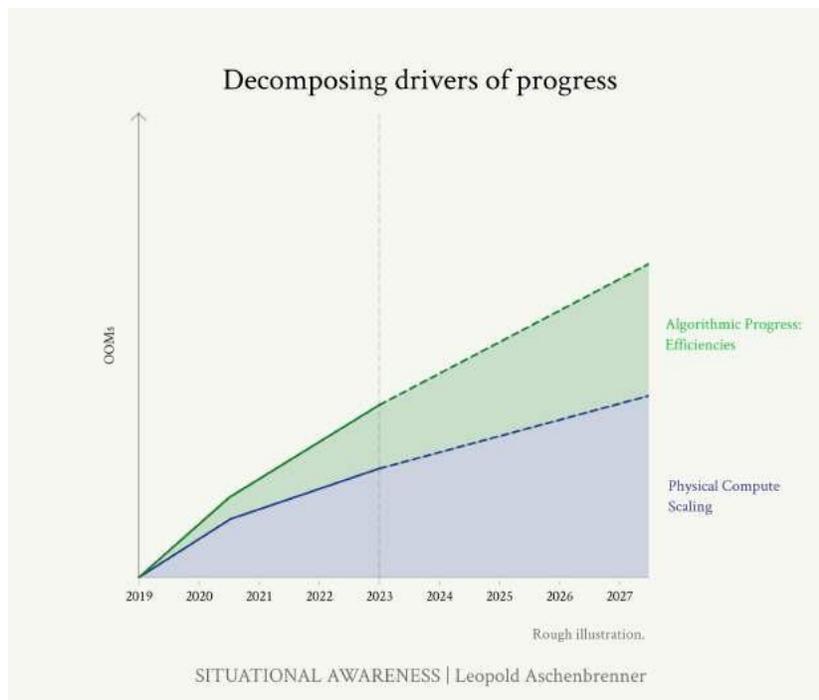
Dicho esto, para ser claros, este es un cálculo muy ingenuo (solo para una ilustración aproximada) que está equivocado de varias maneras. Puede

haber optimizaciones específicas de inferencia (que no se traducen en eficiencia de entrenamiento); puede haber eficiencias de entrenamiento que no reduzcan el recuento de parámetros (y, por lo tanto, no se traduzcan en eficiencia de inferencia); y así sucesivamente.

- Gemini 1.5 Flash, lanzado recientemente, ofrece entre "GPT-Rendimiento de nivel 3,75" y GPT-4,¹⁵ mientras **cálculo de costes** similar a 85x/57x (entrada/salida) menos que el GPT-4 original (¡ganancias extraordinarias!).

- **Leyes de escalado de la chinchilla** dar un 3x+ (0.5 OOMs+) de ganancia de eficiencia.¹⁶
- Gemini 1.5 Pro reclamó **importantes ganancias de eficiencia de cómputo** (superando a Gemini 1.0 Ultra, mientras usaba "significativamente menos" cómputo), con Mix of Experts (MoE) como un cambio de arquitectura destacado. **Otros documentos también afirman un múltiplo sustancial en el cómputo del Ministerio de Educación.**
- Ha habido muchos ajustes y ganancias en **la arquitectura, los datos, la pila de entrenamiento, etc.** Todo el rato.¹⁷

En conjunto, la información pública sugiere que el salto de GPT-2 a GPT-4 incluyó 1-2 OOM de ganancias de eficiencia algorítmica.¹⁸



¹⁵ Gemini 1.5 Flash se clasifica de manera

GPT-4 (más alto que el GPT-4 original, más bajo que las versiones actualizadas de GPT-4) en LMSys, una **tabla de clasificación de chatbots**, y tiene un **rendimiento similar** en MATH y GPQA (evaluaciones que miden el razonamiento) que el GPT-4 original, mientras que aterriza aproximadamente en el medio entre GPT-3.5 y GPT-4 en MMLU (una evaluación que se inclina más hacia la medición del conocimiento).

¹⁶ A escala ~GPT-3, más de 3 veces a escalas mayores.

¹⁷ Por ejemplo, **este artículo** contiene una comparación de un Transformer vainilla estilo GPT-3 con varios cambios simples en la arquitectura y la receta de entrenamiento publicados a lo largo de los años (RMSnorms en lugar de layernorm, diferentes incrustaciones posicionales, activación de SwiGlu, optimizador AdamW en lugar de Adam, etc.), lo que llaman "Transformer++", lo que implica una ganancia de 6x al menos a pequeña escala.

¹⁸ Si tomamos la tendencia de 0,5 OOMs/año, y 4 años entre el lanzamiento de GPT-2 y GPT-4, serían 2 OOMs. Sin embargo, GPT-2 a GPT-3 fue un simple escalado (después de grandes ganancias de, por ejemplo, Transformers), y OpenAI afirma que el **preentrenamiento de GPT-4 terminó en 2022**, lo que podría significar que estamos ante un progreso algorítmico cercano a los 2 años que deberíamos contar aquí. 1 OOM de eficiencia algorítmica parece un límite inferior conservador.

Figura 15: Progreso de la descomposición: eficiencias informáticas y algorítmicas. (Ilustración aproximada.)

A lo largo de los 4 años posteriores a GPT-4, deberíamos esperar la tendencia

Para continuar:¹⁹ en promedio ~0.5 OOMs/año de eficiencia de cómputo

es decir, ~2 OOM de ganancias en comparación con GPT-4

para 2027. Si bien las eficiencias informáticas serán más

difíciles de encontrar a medida que recojamos la fruta

madura, las inversiones de los laboratorios de IA en dinero

y talento para encontrar nuevas mejoras algorítmicas están

creciendo

rápidamente.²⁰ (La inferencia inferible públicamente es rentable,

proceso 3x

al menos, no parece haber disminuido la velocidad en absoluto).

En el extremo superior, incluso podríamos ver avances más

fundamentales, similares a los de Transformer21, con ganancias

aún mayores.

En conjunto, esto sugiere que deberíamos esperar algo así

como 1-3 OOM de ganancias de eficiencia algorítmica (en

comparación con GPT-4) para fines de 2027, tal vez con una

mejor estimación de ~ 2 OOM.

¹⁹ Al menos, dado más de una década de mejoras algorítmicas consistentes, la carga de la prueba recaería en aquellos que sugieran que todo se detendrá repentinamente.

²⁰ Los rendimientos económicos de un

sus modelos sean mejores puede volverse mucho más difícil debido a la restricción de datos. Esto no debe subestimarse: hemos estado montando el

El muro de datos

Hay una fuente potencialmente importante de variación para todo esto: nos estamos quedando sin datos de Internet. Eso podría significar que, muy pronto, el enfoque ingenuo para preentrenar modelos de lenguaje más grandes con más datos extraídos podría comenzar a toparse con serios cuellos de botella.

Los modelos de frontera ya están entrenados en gran parte de la red. Llama 3, por ejemplo, fue entrenado con **tokens de 15T**. Common Crawl, un volcado de gran parte de Internet **utilizado para la formación de LLM**, es >100T de tokens en bruto, aunque gran parte de eso es spam y duplicación (por ejemplo, **una duplicación aráivamente simple** conduce a 30T tokens, lo que implica que Llama 3 ya estaría usando básicamente todos los datos). Además, para dominios más específicos como el código, hay muchos menos tokens todavía, por ejemplo, **se estima que los repositorios públicos de github** están en pocos billones de tokens.

Se puede ir un poco más lejos repitiendo datos, pero **el trabajo académico sobre estos** sugiere que la repetición sólo te lleva hasta cierto punto, encontrando que después de 16 épocas (una repetición de 16 veces), los rendimientos disminuyen extremadamente rápido a cero. En algún momento, incluso con un cómputo más (efectivo), hacer que

La eficiencia se medirá en 10 dólares de miles de millones o más, dados los costos de los clústeres.

²¹Más o menos algo así como una ~ ganancia de 10x.

escalando curvas, montando la ola del paradigma del modelado-lenguaje-preentrenamiento, y sin algo nuevo aquí, este paradigma se agotará (al menos ingenuamente). A pesar de las inversiones masivas, nos estancaríamos.

Se rumorea que todos los laboratorios están haciendo apuestas masivas de investigación sobre nuevas mejoras algorítmicas o enfoques para evitar esto. Supuestamente, los investigadores están probando muchas estrategias, desde **datos sintéticos hasta autojuego y enfoques de RL**. Los conocedores de la industria parecen ser muy optimistas: Dario Amodei (CEO de Anthropic) **dijo recientemente en un podcast**: "si lo miras muy ingenuamente, no estamos tan lejos de quedarnos sin datos [. . .] Mi conjetura es que esto no será un obstáculo [. . .] Hay muchas maneras diferentes de hacerlo". Por supuesto, los resultados de cualquier investigación sobre esto son patentados y no se publican en estos días.

Además del optimismo interno, creo que hay un fuerte argumento intuitivo de por qué debería ser posible encontrar formas de entrenar modelos con una eficiencia de muestra mucho mejor (mejoras algorítmicas que les permitan aprender más de datos limitados). Considere cómo usted o yo aprenderíamos de un libro de texto de matemáticas realmente denso:

- Lo que hace un LLM moderno durante la formación es, esencialmente, hojear muy rápidamente el libro de texto, las palabras simplemente *pasan volando*, sin gastar mucho poder mental en él.
- Más bien, cuando usted o yo leemos ese libro de texto de matemáticas, leemos un par de páginas lentamente; luego tener un monólogo interno sobre el material en nuestras cabezas y hablar de ello con algunos compañeros de estudio; leer una o dos páginas más; luego intentar algunos problemas de práctica, fallar, volver a intentarlo de una manera diferente, obtener comentarios sobre esos problemas, intentarlo de nuevo hasta que obtengamos un problema correcto; y así sucesivamente, hasta que finalmente el material "hace clic".
- Tú o yo tampoco aprenderíamos mucho de un libro de texto de matemáticas denso si todo lo que

pudiéramos hacer fuera
lo atraviesan como LLM.²²
de texto

²²Y simplemente releendo el mismo libro

Una y otra vez puede resultar en la
memorización, no en la comprensión.
¡Supongo que esa es la cantidad de
wordcels que pasan las clases de
matemáticas!

- Pero quizás, entonces, hay formas de incorporar aspectos de cómo los humanos digerirían un libro de texto de matemáticas denso para permitir que los modelos aprendan mucho más de datos limitados. En un sentido simplificado, este tipo de cosas —tener un monólogo interno sobre el material, tener una discusión con un compañero de estudio, probar y fracasar en los problemas hasta que haga clic— es lo que muchos enfoques de datos sintéticos/auto-juego/RL están tratando de hacer.²³

El antiguo estado del arte de los modelos de entrenamiento era simple e ingenuo, pero funcionaba, por lo que nadie se esforzó por descifrar estos enfoques para la eficiencia de la muestra. Ahora que puede convertirse en una restricción, deberíamos esperar que todos los laboratorios inviertan miles de millones de dólares y sus mentes más inteligentes para descifrarlo. Un patrón común en el aprendizaje profundo es que se necesita *mucho* esfuerzo (y muchos proyectos fallidos) para obtener los detalles correctos, pero eventualmente alguna versión de lo obvio y simple simplemente funciona. Teniendo en cuenta cómo el aprendizaje profundo ha logrado romper todos los supuestos muros durante la última década, mi caso base es que será similar aquí.

Además, parece posible que descifrar una de estas apuestas algorítmicas como los datos sintéticos pueda *mejorar drásticamente* los modelos. Aquí hay una bomba de intuición. Los modelos fronterizos actuales, como Llama 3, se entrenan en Internet, e Internet es en su mayoría basura, como el comercio electrónico o el SEO o lo que sea. Muchos LLM gastan la gran mayoría de su formación en esta basura, en lugar de en datos de muy alta calidad (por ejemplo, cadenas de razonamiento de personas que trabajan en problemas científicos difíciles). Imagínese si pudiera dedicar el cómputo de nivel GPT-4 a datos de muy alta calidad: podría ser un modelo mucho, mucho más capaz.

Una mirada retrospectiva a AlphaGo, el primer sistema de IA que venció a los campeones del mundo en Go, décadas antes de que se pensara posible, también es útil aquí.²⁴

Karpathy

- En el paso 1, AlphaGo fue entrenado mediante el

²³ Otra forma de verlo me parece interesante: hay un "punto intermedio perdido" entre el preentrenamiento y el Aprendizaje en contexto.

El aprendizaje en contexto es *increíble* (y competitivo con la eficiencia de la muestra humana). Por ejemplo, el artículo de Gemini 1.5 Pro discute dar al modelo materiales instructivos (un libro de texto, un diccionario) sobre Kalamang, un idioma hablado por menos de 200 personas y básicamente no presente en Internet, en contexto, ¡y el modelo aprende a traducir del inglés al Kalamang a nivel humano! En contexto, el modelo es capaz de aprender del libro de texto tan bien como lo haría un humano (y mucho mejor de lo que aprendería simplemente tirando ese libro de texto en el entrenamiento previo).

Cuando un ser humano aprende de un libro de texto, es capaz de destilar su memoria a corto plazo/aprendizajes en memoria a largo plazo/habilidades a largo plazo con la práctica; Sin embargo, no tenemos una forma equivalente de destilar el aprendizaje en contexto "de vuelta a los pesos". Los datos sintéticos, el juego propio, el RL, etc. están tratando de arreglar eso: dejar que el modelo aprenda por sí mismo, luego pensar en él y practicar lo que aprendió, destilando ese aprendizaje en las pesas.

²⁴ Véase también la charla de Andrej

discutiendo esto aquí.

aprendizaje por imitación en juegos de Go humanos expertos. Esto le dio una base.

- En el paso 2, AlphaGo jugó millones de partidas contra sí mismo. Esto hizo que se volviera sobrehumano en Go: recordemos la famosa jugada 37 en la partida contra Lee Sedol, una jugada extremadamente inusual pero brillante que un humano nunca habría jugado.

Desarrollar el equivalente del paso 2 para los LLM es un problema clave de investigación para superar el muro de datos (y, además, en última instancia, será la clave para superar la inteligencia a nivel humano).

Todo esto quiere decir que las limitaciones de datos parecen inyectar grandes barras de error en cualquier sentido en la predicción de los próximos años del progreso de la IA. Hay una posibilidad muy real de que las cosas se estanquen (los LLM podrían seguir siendo tan importantes como Internet, pero no llegaríamos a un AGI realmente loco). Pero creo que es razonable suponer que los laboratorios lo descifrarán, y que hacerlo no solo mantendrá las curvas de escalado, sino que posiblemente permitirá enormes ganancias en la capacidad del modelo.

Por otro lado, esto también significa que deberíamos esperar más variación entre los diferentes laboratorios en los próximos años en comparación con la actualidad. Hasta hace poco, se publicaban las técnicas más avanzadas, por lo que todo el mundo hacía básicamente lo mismo. (Y los nuevos advenedizos o proyectos de código abierto podrían competir fácilmente con la frontera, ya que se publicó la receta). Ahora, las ideas algorítmicas clave se están volviendo cada vez más propietarias. Esperaría los enfoques de los laboratorios divergir mucho más, y algunos progresar más rápido que otros, incluso un laboratorio que ahora parece estar en la frontera podría quedarse atascado en la pared de datos mientras otros hacen un avance que les permite avanzar. Y el código abierto tendrá muchas más dificultades para competir. Sin duda, hará que las cosas sean interesantes. (Y si y cuando un laboratorio figura Su avance será la clave de AGI, la clave de la superinteligencia, uno de los secretos más preciados de Estados Unidos).

Destrajeando

Finalmente, la categoría de mejoras más difícil de cuantificar, pero no menos importante: lo que llamaré "descojear".

Imagínese si cuando se le pide que resuelva un problema matemático difícil, tuviera que responder instantáneamente con lo primero que se le ocurra. Parece obvio que lo pasarías mal, excepto por los problemas más simples. Pero hasta hace poco, así es como los LLM resolvían problemas matemáticos. En cambio, la mayoría de nosotros trabajamos en el problema paso a paso en un bloc de notas, y somos capaces de resolver problemas mucho más difíciles de esa manera. La "cadena de pensamiento" que impulsó eso desbloqueó eso para los LLM. A pesar de sus excelentes capacidades en bruto, eran mucho peores en matemáticas de lo que podrían ser porque estaban cojeando de una manera obvia, y se necesitó un pequeño ajuste algorítmico para desbloquear capacidades mucho mayores.

En los últimos años, hemos hecho grandes avances en los modelos "descojear". Se trata de mejoras algorítmicas que van más allá del simple entrenamiento de mejores modelos base (y a menudo solo usan una fracción del proceso de preentrenamiento) que liberan las capacidades del modelo:

- *Aprendizaje por refuerzo a partir de la retroalimentación humana (RLHF)*. Mod-
Tienen una experiencia increíble *latente Capacidades*²⁵ pero son crudos y ausencia de supervisión ²⁵ Esa es la magia de la
Es increíblemente difícil trabajar con él. Si bien la concepción popular de RLHF es que simplemente censura las palabrotas, RLHF ha sido clave para hacer que los modelos sean realmente útiles y comercialmente valiosos (en lugar de hacer que los modelos predigan texto aleatorio de Internet, hacer que realmente apliquen sus capacidades para tratar de responder a tu pregunta!). Esta fue la magia de ChatGPT: RLHF bien hecho hizo que los modelos fueran utilizables y útiles para personas reales por primera vez. [El artículo original de InstructGPT](#) tiene una gran cuantificación de esto: un modelo pequeño con RLHF era equivalente a un modelo sin RLHF >100 veces más grande en términos de preferencia de evaluador humano.
- *Andamios*. Piensa en CoT++: en lugar de preguntarle a un modelo
- *Cadena de Pensamiento (CoT)*. Como se ha comentado, [CoT comenzó a usarse ampliamente hace solo 2 años](#) y puede proporcionar el equivalente a un aumento de cómputo efectivo de > 10 veces en problemas matemáticos / de razonamiento.

Aprendizaje, en cierto sentido: para predecir mejor la siguiente ficha, para hacer que disminuya la perplejidad, los modelos aprenden representaciones internas increíblemente ricas, desde el sentimiento (*famoso*) hasta modelos de mundo complejos. Pero, fuera de la caja, están cojeando: están usando sus representaciones internas poco creíbles simplemente para predecir el próximo token en texto aleatorio de Internet, y en lugar de aplicarlas de la mejor manera para tratar de resolver su problema.

Para resolver un problema, haz que un modelo haga un plan de ataque, que otro proponga un montón de posibles soluciones, que otro lo critique, y así sucesivamente.

Porejemplo, en HumanEval (problemas de codificación), un andamiaje simple permite que GPT-3.5 supere a GPT-4 sin andamiaje. En SWE-Bench (un punto de referencia para resolver tareas de ingeniería de software del mundo real), GPT-4 solo puede resolver ~ 2% correctamente, mientras que con el andamiaje de agente de Devinsalta al 14-23%. (Sin embargo, el desbloqueo de la agencia está solo en su infancia, como discutiré más adelante).

- *Herramientas:* Imagínesi si a los humanos no se les permitiera usar calculadoras o computadoras. Solo estamos al principio, pero Chat-GPT ahora puede usar un navegador web, ejecutar código, etc.
- *Longitud del contexto.* Los modelos han pasado de un contexto de token de 2k (GPT-3) a un contexto de 32k (versión GPT-4) a un contexto de 1M+ (Gemini 1.5 Pro). Esto es un *gran* problema. Un modelo base mucho más pequeño con, digamos, 100 mil tokens de contexto relevante puede superar a un modelo que es mucho más grande pero que solo tiene, digamos, 4 mil datos relevantes de contexto: más contexto es efectivamente un proceso grande

Aumento de la eficiencia.²⁶ En términos más generales, el contexto es clave para desbloquear-

Por ejemplo, muchas aplicaciones de codificación requieren la comprensión de grandes partes de estos modelos.

una base de código para contribuir de manera útil con el nuevo código; o, si está utilizando un modelo para ayudarlo a escribir un documento en el trabajo, realmente necesita el contexto de muchos documentos y conversaciones internas relacionados.

Gemini 1.5 Pro, con su texto de 1M+ tokens, *escapaz* de aprender un nuevo idioma (un idioma de bajos recursos que no está en Internet) desde cero, ¡simplemente poniendo en contexto un diccionario y materiales de referencia gramatical!

- *Mejoras posteriores al entrenamiento.* El GPT-4 actual ha mejorado sustancialmente en comparación con el GPT-4 original cuando se vuelve a liberar, según John Schulman debido a las mejoras posteriores al entrenamiento que desbloquearon la capacidad del modelo latente: en las evaluaciones de razonamiento, ha logrado ganancias sustanciales (por ejemplo, ~ 50% -> 72% en MATH, ~ 40% a ~50% en GPQA) y en la tabla de

²⁶ Vea la Figura 7 de la Gema actualizada clasificación de LMSys, ha dado un salto de elo de casi 100 puntos (compara- La diferencia de elo entre el Claude 3 Haiku y el Claude 3 Opus, mucho más grande, son modelos que tienen una diferencia de precio de ~50x).

ini 1.5, comparando la perplejidad frente al contexto para Gemini 1.5 Pro y Gemini 1.5 Flash (un modelo mucho más barato y presumiblemente más pequeño).

Un estudio realizado por Epoch AI sobre algunas de estas técnicas, como el plegado de andamios, el uso de herramientas, etc., revela que técnicas como esta suelen dar lugar a ganancias informáticas efectivas de 5 a 30 veces en muchos puntos de referencia. De manera similar, METR (una organización que evalúa modelos) encontró mejoras de rendimiento muy grandes en su conjunto de tareas agenticas, a través de la eliminación del mismo modelo base de GPT-4: del 5 % con solo el modelo base, al 20 % con el GPT-4 como posentrenado en el lanzamiento, a casi el 40 % actual a partir del mejor post-entrenamiento, herramientas y andamiaje de agentes (Figura 16).

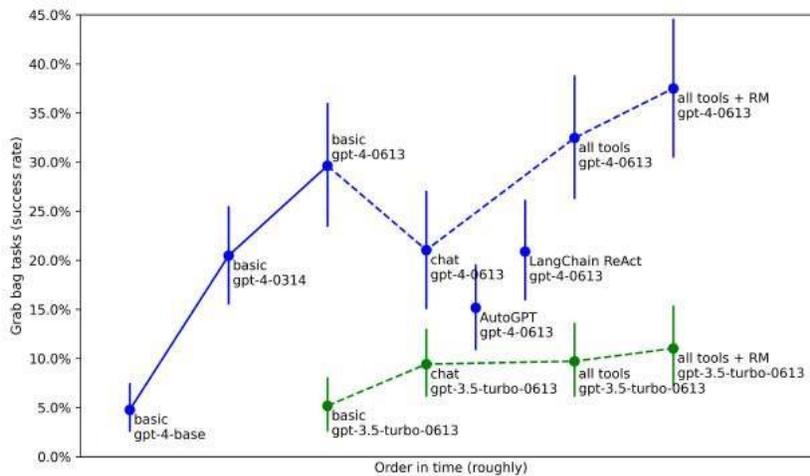


Figura 16: Rendimiento en las tareas agentic de METR, a lo largo del tiempo a través de una mejor liberación. Fuente: Evaluación de modelose investigación de amenazas

Si bien es difícil ponerlos en una escala de cómputo efectiva unificada con eficiencias informáticas y algorítmicas, está claro que se trata de *grandes* ganancias, al menos en una magnitud aproximadamente similar a la de la Escalado de cómputo y eficiencias algorítmicas. (También destaca el papel central del progreso algorítmico: los ~0,5 OOMs/año de las eficiencias de cómputo, ya significativas, son solo una parte de la historia, y junto con el progreso algorítmico sin trabas en general es tal vez incluso una de las ganancias en la tendencia actual).

"Destruir" es una gran parte de lo que realmente permitió que estos modelos se volvieran útiles, y yo diría que gran parte de lo que está frenando muchas aplicaciones comerciales hoy en día es la necesidad de más "destrabas" de este tipo. De hecho, *las modelos de hoy en día siguen cojeando increíblemente*. Por ejemplo:

- No tienen memoria a largo plazo.

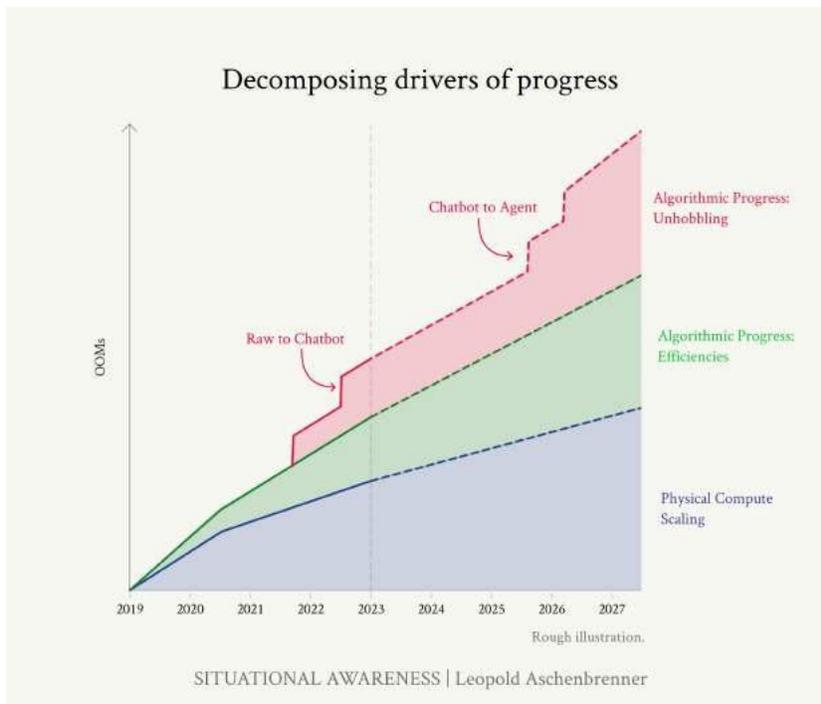


Figura 17: Progreso en descomposición: computación, eficiencias algorítmicas y descompresión. (Ilustración aproximada.)

- No pueden usar una computadora (todavía tienen herramientas muy limitadas).
- Todavía en su mayoría no piensan antes de hablar. Cuando le pides a ChatGPT que escriba un ensayo, es como esperar a un humano para escribir un ensayo a través de su flujo inicial de conciencia.²⁷ en esto.
- Pueden (en su mayoría) participar en breves diálogos de ida y vuelta, en lugar de ausentarse por un día o una semana, pensando en un problema, investigando diferentes enfoques, consultando a otros seres humanos y luego escribiéndote un informe más largo o una solicitud de extracción.
- En su mayoría, no están personalizados para usted o su aplicación (solo un chatbot genérico con un breve mensaje, en lugar de tener toda la información relevante sobre su empresa y su trabajo).

²⁷ Las personas son Sin embargo, trabajando

Las posibilidades aquí son enormes, y estamos recogiendo rápidamente la fruta madura aquí. Esto es fundamental: *es completamente erróneo para imaginar "GPT-6 ChatGPT"*. Con el continuo descojeo

progreso, las mejoras serán cambios radicales en comparación con GPT-6 + RLHF. Para 2027, en lugar de un chatbot, tendrás algo que se parecerá más a un agente, a un compañero de trabajo.

De chatbot a agente-compañero de trabajo

¿Cómo podría ser una ambiciosa liberación en los próximos años? De la forma en que lo pienso, hay tres ingredientes clave:

1. Resolviendo EL "problema de incorporación"

GPT-4 tiene la inteligencia bruta para hacer una parte decente del trabajo de muchas personas, pero es algo así como una nueva contratación inteligente que acaba de aparecer hace 5 minutos: no tiene ningún contexto relevante, no ha leído los documentos de la empresa o el historial de Slack ni ha tenido conversaciones con los miembros del equipo, ni ha dedicado tiempo a comprender la base de código interna de la empresa. Un nuevo empleado inteligente no es tan útil 5 minutos después de llegar, ¡pero es bastante útil un mes después! Parece que debería ser posible, por ejemplo, a través de un contexto muy largo, "incorporar" modelos como lo haríamos con un nuevo compañero de trabajo humano. Esto por sí solo sería un gran desbloqueo.

2. EL voladizo DE CÁLCULO en tiempo ST

(RAZONAMIENTO/CORRECCIÓN DE errores/SISTEMA II PARA PROBLEMAS DE HORIZONTE MÁS LARGO)

En este momento, los modelos básicamente solo pueden hacer tareas cortas: les haces una pregunta y te dan una respuesta. Pero eso es extremadamente limitante. El trabajo cognitivo más útil que realizan los humanos es un horizonte más largo: no solo toma 5 minutos, sino horas, días, semanas o meses.

Un científico que solo podía pensar en un problema difícil durante 5 minutos no podía hacer ningún avance científico. Un ingeniero de software que solo pudiera escribir código esqueleto para una sola función cuando se le pidiera no sería muy útil.

A los ingenieros de software se les asigna una tarea más grande, y luego van a hacer un plan, comprender partes

relevantes de la base de código o

herramientas técnicas, escribir diferentes módulos y probarlos de manera increíble, depurar errores, buscar en el espacio de posibles soluciones y, finalmente, enviar una gran solicitud de extracción que es la culminación de semanas de trabajo. Y así sucesivamente.

En esencia, hay un gran *desbordamiento de proceso en tiempo de prueba*. Piensa en cada token GPT-4 como una palabra de monólogo interno cuando piensas en un problema. Cada token GPT-4 es bastante inteligente, pero actualmente solo se puede usar de manera realmente efectiva en el orden de ~ cientos de fichas para cadenas de pensamiento coherentes (efectivamente como si solo pudieras dedicar unos minutos de monólogo/pensamiento interno a un problema o proyecto).

¿Qué pasaría si pudiera usar millones de tokens para pensar y trabajar en problemas realmente difíciles o proyectos más grandes?

Número de tokens	Equivalente a que yo trabajara en algo para...	
Años 100	Unos minutos	ChatGPT (estamos aquí)
1.000	Media hora	+1 Cómputo en tiempo de prueba de OOM
10.000	Media jornada laboral	+2 OOM
100.000	Una semana laboral	+3 OOM
Millones	Varios meses	+4 OOM

Incluso si la inteligencia "por token" fuera la misma, sería la diferencia entre que una persona inteligente gastara unos *pocos minutos* frente a unos *pocos meses* en un problema. No sé ustedes, pero hay mucho, mucho, mucho más de lo que soy capaz de en unos pocos meses frente a unos pocos minutos. Si pudiéramos desbloquear "ser capaces de pensar y trabajar en algo durante meses, equivalente, en lugar de unos pocos minutos" para los modelos, se desbloquearía un *salto loco* en la capacidad. Hay un gran voladizo aquí, que vale la pena muchos OOM.

En este momento, los modelos aún no pueden hacer esto. Incluso con los avances recientes en el contexto largo, este contexto más largo en su mayoría solo funciona para el

consumo de tokens, no para la producción de tokens: después de un tiempo, el modelo se descarrila o se atasca. Todavía no puede desaparecer por un tiempo para trabajar en un

Tabla 2: Suponiendo que un humano piensa a ~ 100 tokens/minuto y trabaja 40 horas/semana, traduciendo "cuánto tiempo un modelo piensa" en fichas al tiempo humano en un problema/proyecto dado.

problema o proyecto por sí solo.²⁸

Pero desbloquear el cómputo en tiempo de prueba podría ser simplemente una cuestión de victorias algorítmicas relativamente pequeñas de "destraba". Tal vez una pequeña cantidad de RL ayuda a un modelo a aprender a corregir errores ("hm, eso no se ve bien, déjame verificarlo"), hacer planes, buscar posibles soluciones, etc. En cierto sentido, el modelo ya tiene la mayoría de las capacidades en bruto, solo necesita aprender algunas habilidades adicionales en la parte superior para ponerlo todo junto.

En esencia, solo necesitamos enseñarle al modelo una especie de Sistema

II bucle exterior²⁹ que le permite razonar a través de Proyectos de horizonte.

Si logramos enseñar este bucle externo, en lugar de una breve respuesta de chatbot de un par de párrafos, imagina un flujo de millones de palabras (que llegan más rápido de lo que puedes leerlas) mientras el modelo piensa en problemas, usa herramientas, prueba diferentes enfoques, investiga, revisa su trabajo, se coordina con otros y completa grandes proyectos por su cuenta.

Intercambio de tiempo de prueba y tiempo de entrenamiento de cómputo en otros dispositivos de ML. En otros dominios, como los sistemas de IA para juegos de mesa, se ha demostrado que se puede usar más proceso en tiempo de prueba (también llamado proceso en tiempo de inferencia) para sustituir el proceso de entrenamiento.

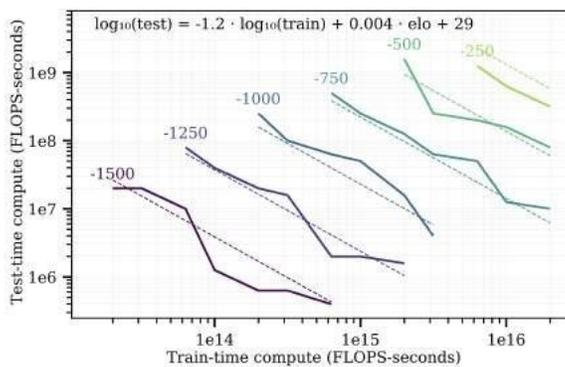


Fig. 9. The trade-off between train-time compute and test-time compute. Each dotted line gives the minimum train-test compute required for a certain Elo on a 9×9 board

²⁸ Lo cual tiene sentido: ¿por qué habría aprendido las habilidades durante más tiempo?

¿Razonamiento de horizonte y corrección de errores?

Hay muy pocos datos en Internet en forma de "mi monólogo interno completo, razonamiento, todos los pasos relevantes en el transcurso de un mes como Trabajo en un proyecto". Desbloquear esta capacidad requerirá un nuevo tipo de entrenamiento, para que aprenda estas habilidades adicionales.

O como dijo Gwern (corresponsal privada): "Cerebro del tamaño de una galaxia, ¿y qué me piden que haga? ¡Predice las respuestas mal escritas en los puntos de referencia! Marvin, la red neuronal deprimida, gimió.

²⁹ El Sistema I vs. el Sistema II es una forma de pensar sobre las capacidades actuales de los LLM, incluidas sus limitaciones y errores tontos, y lo que podría ser posible con RL y la liberación. Piénsalo de esta manera: cuando estás conduciendo, la mayor parte del tiempo estás en piloto automático (sistema I, lo que más hacen los modelos en este momento). Pero cuando te encuentres con una zona de construcción compleja o una intersección novedosa, puedes pedirle a tu pasajero-asiento-acompañante que haga una pausa en tu conversación por un momento mientras averiguas, realmente piensas, qué está sucediendo y qué hacer. Si te vieras obligado a vivir solo con el sistema I (más cercano a los modelos actuales), tendrías muchos problemas. La creación de la capacidad de los bucles de razonamiento del sistema II es un desbloqueo central.

Figura 18: Jones (2021): Un modelo más pequeño puede funcionar tan bien como un modelo mucho más grande en el juego de Hex si le das más tiempo de cálculo de prueba ("más tiempo para pensar"). En este dominio, encuentran que uno puede gastar ~ 1.2 OOM más calcule en el momento de la prueba para obtener un rendimiento equivalente al de un modelo con ~1 OOM más de proceso de entrenamiento.

Si una relación similar se mantuviera en nuestro caso, si pudiéramos desbloquear +4 OOM de cálculo en tiempo de prueba, que podrían ser equivalentes a +3 OOM de cómputo previo al entrenamiento, es decir, algo muy parecido al salto entre GPT-3 y GPT-4. (Resolver este "descojear" equivaldría a una gran ampliación de OOM).

3. USAR UNA computadora

Este es quizás el más sencillo de los tres. Chat: GPT en este momento es básicamente como un humano que se sienta en un cuadro aislado al que puedes enviar mensajes de texto. Si bien las primeras mejoras sin trabas enseñan a los modelos a utilizar herramientas aisladas individuales, Espero que con los modelos multimodales pronto seamos capaces de hacer esto de una sola vez: simplemente permitiremos que los modelos usen una computadora como lo haría un humano.

Eso significa unirse a sus llamadas de Zoom, investigar cosas en línea, enviar mensajes y correos electrónicos a las personas, leer documentos compartidos, usar sus aplicaciones y herramientas de desarrollo, etc. (Por supuesto, Para que los modelos aprovechen al máximo esto en bucles de horizonte más largo, esto irá de la mano con el desbloqueo del cálculo en tiempo de prueba).

A PARTIRde AHORA, espero que obtengamos algo que se parezca mucho a un *trabajador remoto sin cita previa*. Un agente que se une a su empresa, se incorpora como un nuevo empleado humano, le envía mensajes a usted y a sus colegas en Slack y usa sus softwares, realiza solicitudes de incorporación de cambios y eso, dados los grandes proyectos, puede hacer el equivalente al modelo de un humano que se va durante semanas para completar el proyecto de forma independiente. Probablemente necesitará modelos base algo mejores que GPT-4 para desbloquear esto, pero posiblemente ni siquiera *mucho* mejor: hay mucho jugo en arreglar las formas claras y básicas en que los modelos aún están cojeando.

(Un primer vistazo a cómo podría ser esto es [Devin](#), un prototipo temprano de desbloquear el "saliente de agencia" / "voladizo de cómputo en tiempo de prueba" en modelos en el camino hacia la creación de un ingeniero de

software totalmente automatizado. No sé qué tan bien funciona Devin en la práctica, y esta demostración sigue siendo muy

Limitado en comparación con lo que produciría un chatbot adecuado → agente sin cojear, pero es un adelanto útil del tipo de cosas que vendrán pronto).

Por cierto, la centralidad de la liberación de las cojeras podría conducir a un efecto de "boom sónico" algo interesante en términos de aplicaciones comerciales. Los modelos intermedios entre ahora y el trabajador remoto requerirán toneladas de trabajo para cambiar los flujos de trabajo y construir una infraestructura para integrarse y obtener valor económico. El trabajador remoto sin cita previa será mucho más fácil de integrar, solo que, bueno, colóquelo para automatizar todo los trabajos que se podrían realizar de forma remota. Parece plausible que el schlep lleve más tiempo que el deslastramiento, es decir, para cuando el trabajador remoto sea capaz de automatizar un gran número de trabajos, los modelos intermedios aún no se habrán aprovechado e integrado por completo, por lo que el salto en el valor económico generado podría ser algo discontinuo.

Los próximos cuatro años

Juntando LOS números, deberíamos esperar (aproximadamente) otro salto del tamaño de GPT-2 a GPT-4 en los 4 años siguientes a GPT-4, para finales de 2027.

- GPT-2 a GPT-4 fue aproximadamente un escalado de cómputo efectivo base de 4,5 a 6 OOM (eficiencia de cómputo físico y algorítmica), además de importantes ganancias de "desaceleración" (desde el modelo base hasta el chatbot).
- En los 4 años siguientes, deberíamos esperar de 3 a 6 OOM de escalado de cómputo efectivo base (eficiencias algómicas de cómputo físico), con quizás una mejor estimación de ~ 5 OOM, además de cambios graduales en la utilidad y las aplicaciones desbloqueadas por "desvinculación" (de chatbot a agente/trabajador remoto sin contacto).

Para poner esto en perspectiva, supongamos que el entrenamiento de GPT-4 duró 3 meses. *En 2027, un laboratorio líder de IA podrá entrenar un modelo de nivel GPT-4 en un minuto.*³⁰ El escalado vertical de proceso efectivo de OOM

³⁰ Sobre la base de las suposiciones de mejor estimación sobre el cálculo físico y la eficiencia algorítmica de las escalas descritas anteriormente, y las consideraciones de simplificación del paralelismo (en realidad, podría parecerse más a "1440 (60*24) modelos de nivel GPT-4 en un día" o similar).

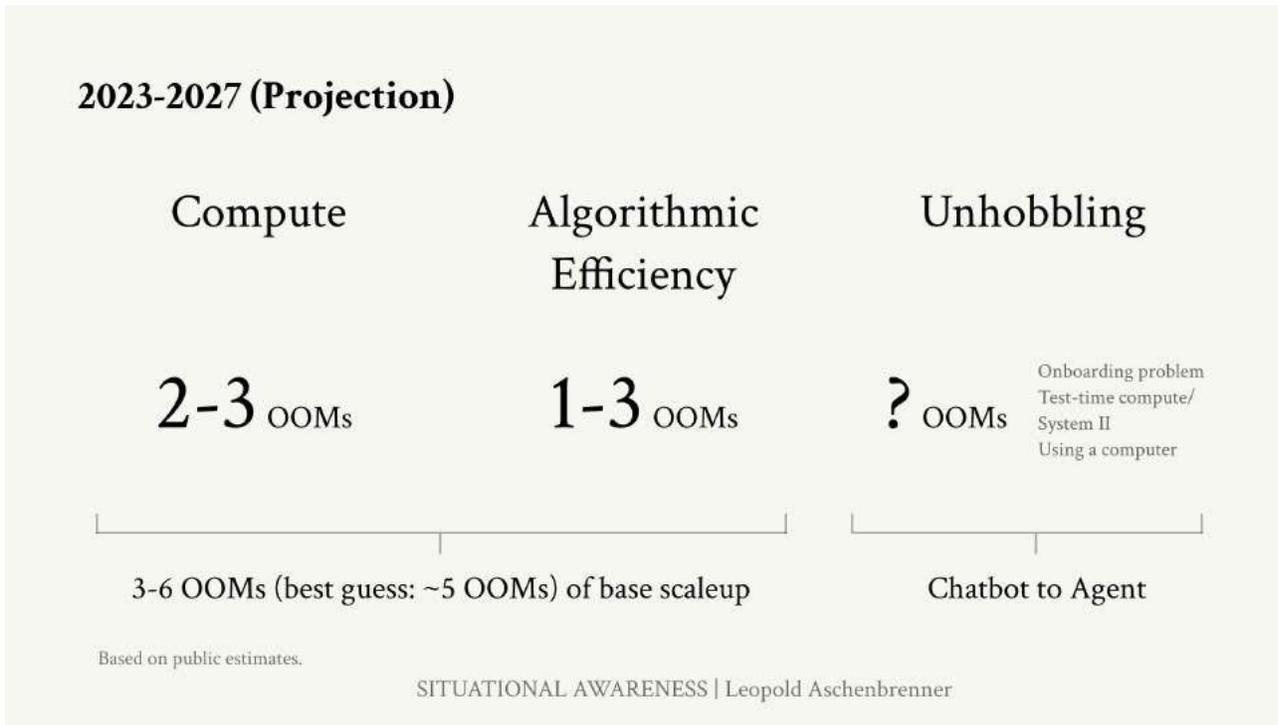
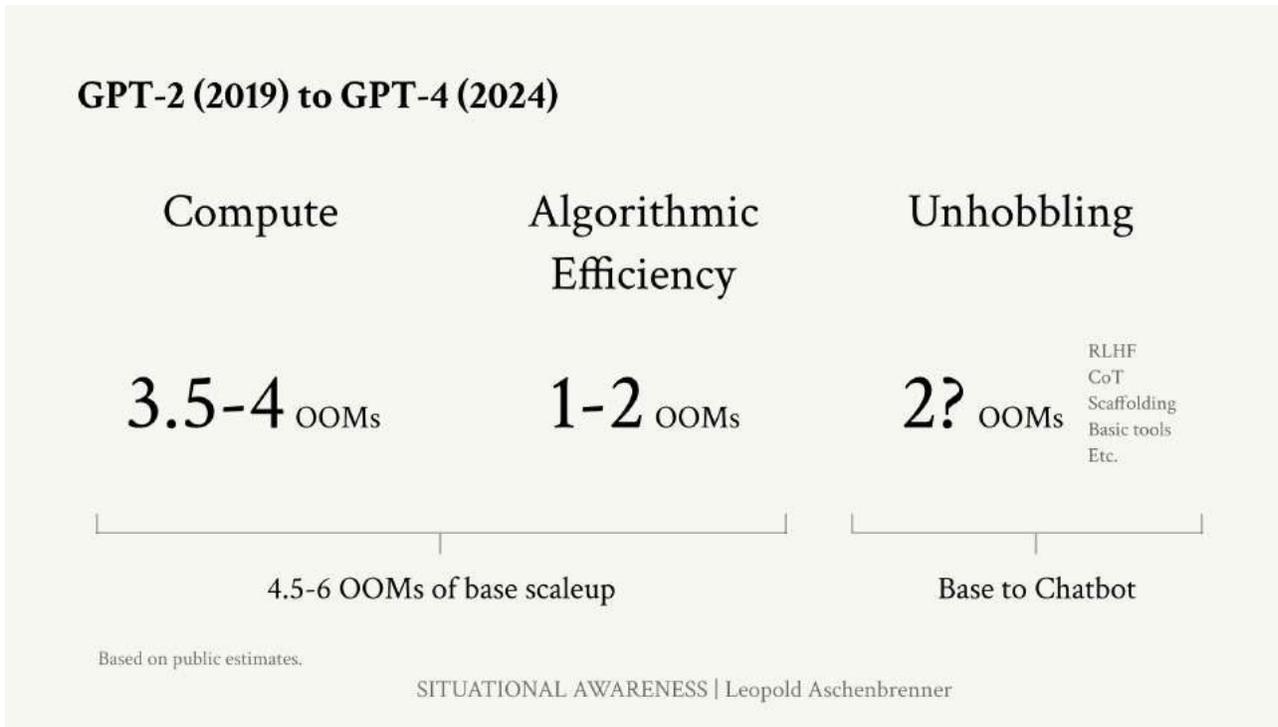


Gráfico 19: Resumen de las estimaciones sobre los impulsores del progreso en los cuatro años anteriores a GPT-4 y lo que deberíamos esperar en los cuatro años siguientes a GPT-4.

será dramático.

¿A dónde nos llevará eso?

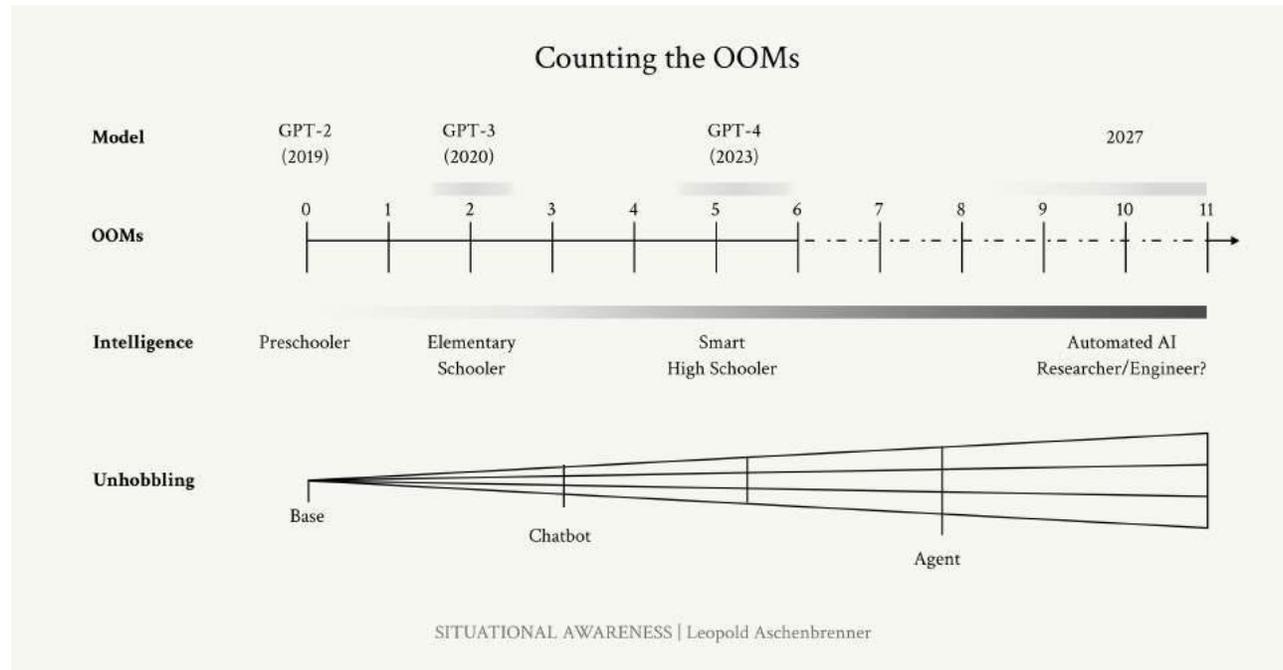


Figura 20: Resumen del recuento de las OOM.

De GPT-2 a GPT-4 nos llevó de ~preescolar a ~inteligente estudiante de secundaria; de apenas ser capaz de producir algunas sentencias cohesivas a superar los exámenes de la escuela secundaria y ser un asistente de codificación útil. Fue un salto de locura. Si *esta* es la inteligencia

Brecha que cubriremos una vez más, ¿a dónde nos llevará eso?³¹ Nosotros ³¹ Por supuesto, cualquier punto de referencia que tengamos

No debería sorprendernos si eso nos lleva muy, muy lejos. Probable

nos llevará a modelos que pueden superar a los doctores y a los mejores expertos en un campo.

Hoy estará saturado. Pero eso no es decir mucho; Es sobre todo una reflexión sobre la dificultad de hacer puntos de referencia lo suficientemente duros.

(Una buena manera de pensar en esto es que la tendencia actual del progreso de la IA avanza aproximadamente 3 veces más rápido que el ritmo de desarrollo infantil. Su hijo de 3 veces la velocidad acaba de graduarse de la escuela secundaria; ¡Te quitará el trabajo antes de que te des cuenta!)

Una vez más, críticamente, no se limite a imaginar un Chat-GPT increíblemente inteligente: las ganancias sin trabas deberían significar que esto se parece más a un trabajador remoto sin cita previa, un agente increíblemente inteligente que puede razonar, planificar y corregir errores y sabe todo sobre usted y su empresa y puede trabajar en un problema independiente.

De hecho, la mayoría de las personas que se encuentran en el centro de la

Estamos en camino de lograr AGI para 2027. Estos sistemas de IA serán básicamente capaces de automatizar básicamente todos los trabajos cognitivos (piense: todos los trabajos que podrían realizarse de forma remota).

Para ser claros, las barras de error son grandes. El progreso podría estancarse a medida que nos quedemos sin datos, si los avances algorítmicos necesarios para atravesar el muro de datos resultan más difíciles de lo esperado.

Tal vez la liberación no llegue tan lejos, y nos quedemos atascados con chatbots meramente expertos, en lugar de compañeros de trabajo expertos. Tal vez las líneas de tendencia de una década se rompan, o la ampliación del aprendizaje profundo se tope con un muro real esta vez. (O un avance algorítmico, incluso un simple desbloqueo que desata el exceso de cómputo en tiempo de prueba, podría ser un cambio de paradigma, acelerando aún más las cosas y conduciendo a AGI incluso antes).

En cualquier caso, estamos corriendo a través de los OOMs, y no se requieren creencias esotéricas, simplemente la extrapolación de tendencias de líneas rectas, para tomar muy en serio la posibilidad de AGI —verdadero AGI— para 2027 .

Parece que muchos están en el juego de definir a la baja el AGI en estos días, como un chatbot realmente bueno o lo que sea. Qué

Quiero decir que es un sistema de IA que podría automatizar completamente mi trabajo o el de mis amigos, que podría hacer completamente el trabajo de un investigador o ingeniero de IA. Tal vez algunas áreas, como la robótica, podrían tardar más en resolverse de forma predeterminada. Y el despliegue social, por ejemplo, en las profesiones médicas o jurídicas, podría verse fácilmente frenado por las decisiones sociales o la regulación. Pero una vez que los modelos pueden automatizar la IA investigación en sí misma, eso es suficiente, suficiente para iniciar intensos bucles de retroalimentación, y podríamos avanzar muy rápidamente, los propios ingenieros de IA automatizada resolverán todos los cuellos de botella restantes para automatizar completamente todo. En particular, millones de investigadores automatizados podrían comprimir muy plausiblemente una década de progreso algorítmico adicional en un año o menos. AGI no será más que una pequeña muestra de la superinteligencia que pronto le seguirá. (Más sobre esto en el próximo capítulo).

En cualquier caso, no esperes que el ritmo vertiginoso del progreso disminuya. Las líneas de tendencia parecen inocentes, pero sus implicaciones son

intenso. Al igual que con todas las generaciones anteriores, cada nueva generación de modelos dejará estupefactos a la mayoría de los espectadores; se mostrarán incrédulos cuando, muy pronto, los modelos resuelvan problemas científicos increíblemente difíciles que llevarían días a los doctores, cuando estén zumbando alrededor de su computadora haciendo su trabajo, cuando estén escribiendo bases de código con millones de líneas de código desde cero, cuando cada año o dos el valor económico generado por estos modelos sea 10 veces mayor. Olvídate de la ciencia ficción, cuenta los OOMs: es lo que deberíamos esperar. AGI ya no es una fantasía lejana. La ampliación de las técnicas sencillas de aprendizaje profundo acaba de funcionar, los modelos sólo quieren aprender, y estamos a punto de hacer otras 100.000x+ para finales de 2027. No pasará mucho tiempo antes de que sean más inteligentes que nosotros.



Figura 21: GPT-4 es solo el comienzo, ¿dónde estaremos cuatro años después? No cometa el error de subestimar el rápido ritmo del progreso del aprendizaje profundo (como lo ilustra el progreso en las GAN).

Adenda. Corriendo a través de los OOMs: Es esta década o fracaso

Solía ser más escéptico con respecto a los plazos cortos de AGI. Una de las razones es que no parecía razonable privilegiar esta década, concentrando tanta masa de probabilidad de AGI en ella (parecía una falacia clásica pensar "oh, somos tan especiales"). Pensé que no debíamos estar seguros de lo que se necesita para obtener AGI, lo que debería conducir a una distribución de probabilidad mucho más "difuminada" sobre cuándo podríamos obtener AGI.

Sin embargo, he cambiado de opinión: críticamente, nuestra incertidumbre sobre lo que se necesita para obtener AGI debería ser sobre *OOM* (de computación efectiva), en lugar de sobre años.

Estamos compitiendo a través de los OOM de esta década. Incluso en su apogeo pasado, la ley de Moore era de solo 1 a 1,5 OOM/década. Estimo que haremos ~5 OOMs en 4 años, y más ~10 en esta década en general.

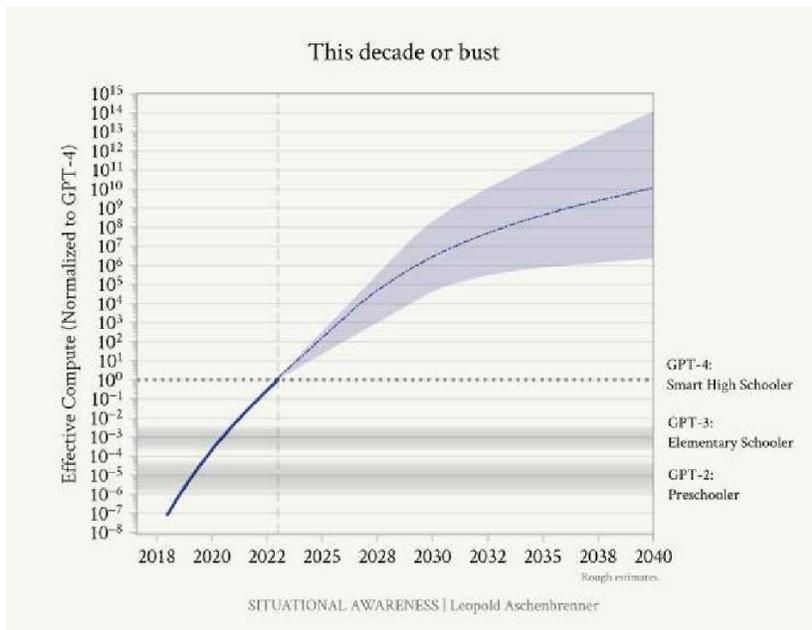


Figura 22: Proyecciones aproximadas sobre el escalado de cómputo eficaz. Hemos estado compitiendo a través de los OOM esta década; Después de principios de la década de 2030, nos enfrentaremos a un trabajo lento.

En esencia, estamos en medio de una enorme expansión que está cosechando ganancias únicas en esta década, y el progreso a través de las OOM será múltiplo más lento a partir de entonces. Si esta ampliación no nos lleva a AGI en los próximos 5 a 10 años, podría ser un

Muy lejos.

- *Ampliación del gasto*: Gastar un millón de dólares en un modelo solía ser escandaloso; para finales de la década, es probable que tengamos grupos de **100.000 millones de dólares o 1.000 millones de dólares**. Ir mucho más alto que eso será difícil; ese ya es básicamente el límite factible (tanto en términos de lo que las grandes empresas pueden permitirse, como incluso solo como una fracción del PIB). A partir de entonces, todo lo que tenemos es una tendencia glacial de crecimiento del PIB real del 2% anual para aumentar esto.
- *Ganancias de hardware*: El hardware de IA ha mejorado mucho más rápido que la ley de Moore. Esto se debe a que hemos estado especializando chips para cargas de trabajo de IA. Por ejemplo, hemos pasado de las CPU a las GPU; chips adaptados para Transformers; y hemos bajado a formatos de números de precisión mucho más bajos, desde fp64/fp32 para la supercomputación tradicional hasta fp8 en H100. Estas son grandes ganancias, pero para finales de la década es probable que tengamos chips específicos de IA totalmente especializados, sin que sea posible obtener mucho más allá de la ley de Moore.
- *Progreso algorítmico*: En la próxima década, los laboratorios de IA invertirán decenas de miles de millones en investigación y desarrollo algorítmico, y todas las personas más inteligentes del mundo trabajarán en esto; desde pequeñas eficiencias hasta nuevos paradigmas, recogeremos muchos de los frutos que están al alcance de la mano. Es probable que no lleguemos a ningún tipo de límite estricto (aunque es probable que los "descojeos" sean finitos), pero al menos el ritmo de las mejoras debería ralentizarse, ya que el rápido crecimiento (en dólares e inversiones en capital humano) necesariamente se ralentiza (por ejemplo, la mayoría de los talentos inteligentes de STEM ya estarán trabajando en IA). (Dicho esto, esto es lo más incierto de predecir, y la fuente de la mayor parte de la incertidumbre sobre los OOM en la década de 2030 en el gráfico anterior).

En conjunto, esto significa que estamos compitiendo a través de muchos más OOM en la próxima década de lo que podríamos hacer en varias décadas posteriores. Tal vez sea suficiente, y pronto obtengamos AGI, o

podríamos tener un trabajo largo y lento. Usted y yo podemos estar razonablemente en desacuerdo sobre el *tiempo medio* hasta el ingreso bruto ajustado,

dependiendo de lo difícil que creamos que será lograr el AGI, pero dado cómo estamos compitiendo a través de los OOM en este momento, ciertamente su *año de AGI modal* debería en algún momento más adelante de esta década más o menos.



Matthew Barnett
@MatthewJBar

...

My own basic calculations suggest that, given the potential for increased investment and hardware progress, we could very soon move through a large fraction of the remaining compute gap between the current frontier models and the literal amount of computation used by evolution.



7:54 PM · Mar 26, 2024 · 3,968 Views

Figura 23: Matthew Barnett tiene una buena visualización relacionada de esto, considerando solo los límites de cómputo y biológicos.

II. De la AGI a la Superinteligencia: la Explosión de Inteligencia

El progreso de la IA no se detendrá a nivel humano. Cientos de millones de AGI podrían automatizar la investigación de IA, comprimiendo una década de progreso algorítmico (5+ OOMs) en 1 año. Pasaríamos rápidamente del nivel humano a sistemas de IA enormemente sobrehumanos. El poder —y el peligro— de la superinteligencia sería dramático.

Definamos una máquina ultrainteligente como una máquina que puede superar con creces todas las actividades intelectuales de cualquier hombre, por muy inteligente que sea. Dado que el diseño de máquinas es una de estas actividades intelectuales, una máquina ultrainteligente podría diseñar máquinas aún mejores; Entonces incuestionablemente habría una "explosión de inteligencia", y la inteligencia del hombre quedaría muy atrás. Así, la primera máquina ultrainteligente es el último invento que el hombre necesita hacer.

i. j. BUENO (1965)

La Bomba y el Súper

En el imaginario común, los terrores de la Guerra Fría se remontan principalmente a Los Álamos, con la invención de la bomba atómica. Pero The Bomb, por sí sola, está quizás sobrevalorada. Podría decirse que pasar de La Bomba a La Súper, bombas de hidrógeno, fue igual de importante.

En los ataques aéreos de Tokio, cientos de bombarderos lanzaron miles de toneladas de bombas convencionales sobre la ciudad. Más tarde ese mismo año, Little Boy, lanzado sobre Hiroshima, desató un poder destructivo similar en un solo dispositivo. Pero solo 7 años después, la bomba de hidrógeno de Teller multiplicada *por mil una vez* más, una sola bomba con más poder explosivo que todas las bombas lanzadas en toda la Segunda Guerra Mundial combinadas.

La Bomba fue una campaña de bombardeo más eficiente. El Súper era un Dispositivo aniquilador de países.³²
Guerra Fría

Lo mismo ocurrirá con AGI y Superinteligencia.

³²Y gran parte de la persecución de la

versidades (cf. [el libro de Daniel Ellsberg](#))
otroGPT- 2-to-

El progreso de la IA NO se detendrá EN EL NIVEL DE HUMAN. Después de aprender inicialmente de los mejores juegos humanos, AlphaGo comenzó a jugar contra sí mismo, y rápidamente se convirtió en sobrehumano, jugando movimientos extremadamente creativos y complejos que a un humano nunca se le habrían ocurrido.

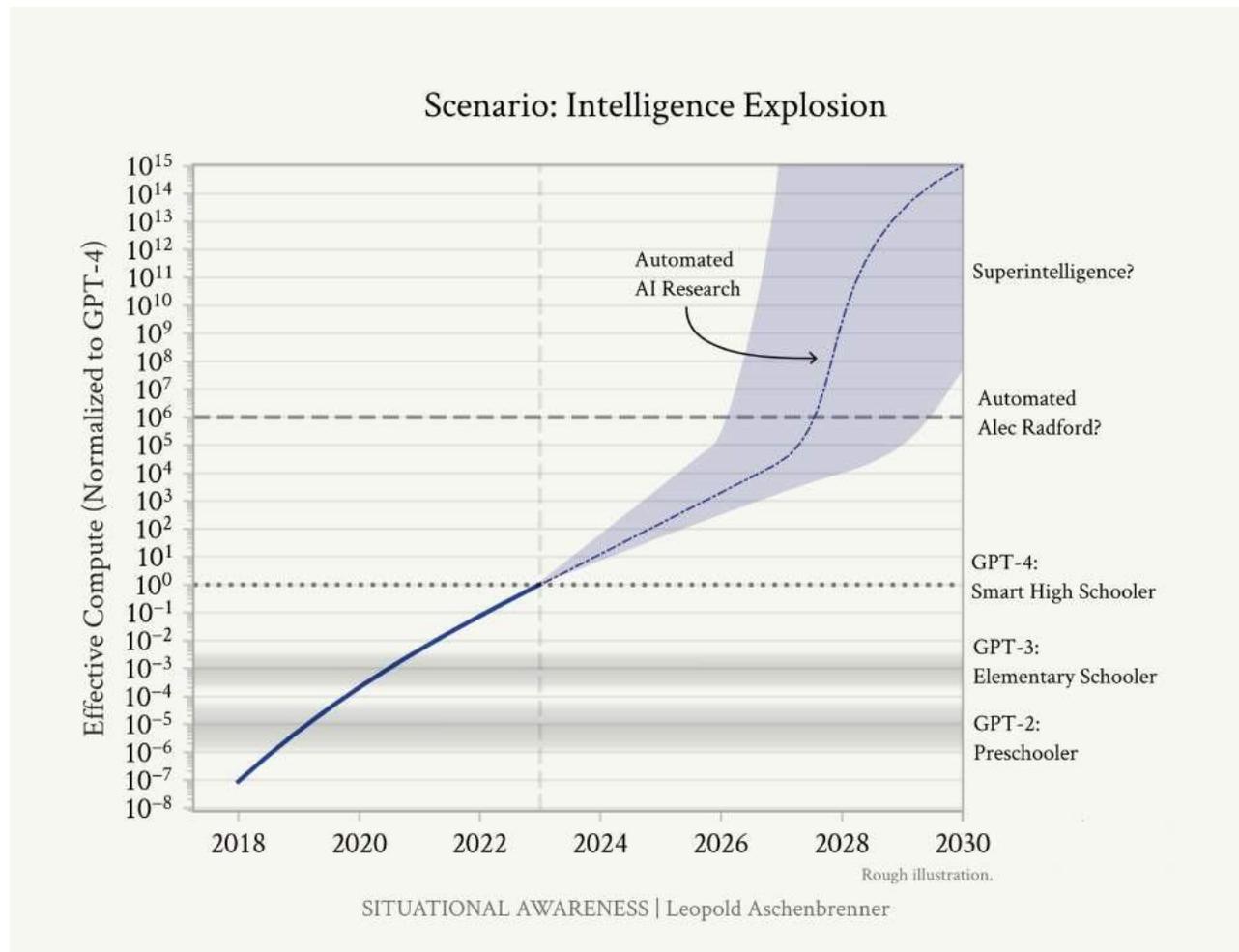
Discutimos el camino hacia la AGI en el artículo anterior. Una vez que obtengamos AGI, giraremos la manivela una vez más, o dos o tres veces más, y los sistemas de IA se volverán sobrehumanos, enormemente sobrehumanos. Llegarán a ser cualitativamente más inteligentes que tú o que yo, mucho más inteligentes, tal vez de manera similar a como tú o yo somos cualitativamente más inteligentes que un niño de primaria.

El salto a la superinteligencia sería bastante salvaje al ritmo rápido pero continuo actual del progreso de la IA (si pudiéramos dar el salto a AGI en 4 años desde GPT-4, ¿qué podrían traer otros 4 u 8 años después de eso?). Pero podría ser mucho más rápido que eso, si AGI automatiza la investigación de IA por sí misma.

Una vez que obtengamos AGI, no solo tendremos un AGI. Revisaré los números más adelante, pero: dada la inferencia de las flotas de GPU para entonces, Es probable que podamos ejecutar *muchos millones de ellos (tal vez 100 millones equivalentes humanos, y poco después a una velocidad humana de 10x+)*. Incluso si aún no pueden caminar por la oficina o preparar café, podrán hacer una investigación de ML en una computadora. En lugar de unos pocos cientos de investigadores e ingenieros en un laboratorio líder de IA, tendríamos más de 100.000 veces eso, trabajando furiosamente en avances algorítmicos, día y noche. Sí, recursivo superación personal, pero no se requiere ciencia ficción; *sólo tendrían que acelerar las líneas de tendencia existentes del progreso algorítmico (actualmente en ~0.5 OOMs/año)*.

La investigación automatizada de la IA probablemente podría comprimir una década humana de progreso algorítmico en menos de un año (y eso parece conservador). Eso sería 5+ OOMs,

se derivaron de la mera sustitución de las bombas A por bombas H, sin ajustar la política nuclear y los planes de guerra al aumento masivo de la capacidad.



Salto del tamaño de GPT-4, por encima de AGI, un salto cualitativo como el de un niño en edad preescolar a un estudiante de secundaria inteligente, *por encima* de los sistemas de IA que ya son tan inteligentes como los investigadores/ingenieros expertos en IA.

Hay varios cuellos de botella plausibles, incluida la computación limitada para los experimentos, las complementariedades con los humanos y el progreso algorítmico cada vez más difícil, que abordaré, pero ninguno parece suficiente para ralentizar definitivamente las cosas.

Antes de que nos demos cuenta, tendríamos superinteligencia en nuestras manos: sistemas de IA *mucho* más inteligentes que los humanos, capaces de un comportamiento novedoso, creativo y complicado que ni siquiera podríamos comenzar a entender, tal vez incluso una pequeña civilización de miles de millones de ellos. Su poder también sería enorme. Aplicación de superin-

Figura 24: La investigación automatizada de IA podría acelerar el progreso algorítmico, lo que llevaría a 5+ OOM de ganancias informáticas efectivas en un año. Los sistemas de IA que tendríamos Al final de una explosión de inteligencia, sería mucho más inteligente que los humanos.

De acuerdo con la investigación y el desarrollo en otros campos, el progreso explosivo se ampliaría a partir de la investigación de ML; Pronto resolverían la robótica, darían saltos dramáticos en otros campos de la ciencia y la tecnología en cuestión de años, y se produciría una explosión industrial. Es probable que la superinteligencia proporcionara una ventaja militar decisiva y desplegara incalculables poderes de destrucción. Nos enfrentaremos a uno de los momentos más intensos y volátiles de la historia de la humanidad.

Automatización de la investigación en IA

No necesitamos automatizarlo todo, solo la investigación de IA. Una objeción común a los impactos transformadores de la AGI es que será difícil para la IA hacer todo. Fíjate en la robótica, por ejemplo, dicen los escépticos; eso será un problema complicado, incluso si la IA está cognitivamente al nivel de los doctorados. O tomemos como ejemplo la automatización de la investigación y el desarrollo de la biología, que podría requerir mucho trabajo físico de laboratorio y experimentos en humanos.

Pero no necesitamos robótica, no necesitamos muchas cosas, para que la IA automatice la investigación de la IA. El trabajo de los investigadores e ingenieros de IA en los principales laboratorios se puede realizar de forma totalmente virtual y no se encuentra con cuellos de botella en el mundo real de la misma manera (aunque seguirá estando limitado por la computación, que abordaré más adelante). Y el trabajo de un investigador de IA es bastante sencillo, en el gran esquema de las cosas: leer literatura de ML y proponer nuevas preguntas o ideas, implementar experimentos para probar esas ideas, interpretar los resultados y repetir. Todo esto parece estar en el ámbito en el que las simples extrapolaciones de las capacidades actuales de la IA podrían llevarnos fácilmente a los niveles de los mejores humanos o más allá de ellos para finales de 2027.³³

de retroalimentación
extraordinarios.³⁴

Vale la pena enfatizar lo sencillos y complicados que han sido algunos de los mayores avances en aprendizaje automático de la última década: "oh, solo agregue algo de normalización" (LayerNorm / BatchNorm) o "haga $f(x)+x$ en lugar de $f(x)$ " (conexiones residuales) o "corrija un error de implementación" (leyes de escalado de Kaplan → Chinchilla). La investigación en IA puede automatizarse. Y la investigación de la IA es todo lo que se necesita para poner en marcha bucles

³³El trabajo de un investigador de IA también es un trabajo que los investigadores de IA en los laboratorios de IA conocen muy bien, por lo que será particularmente intuitivo para ellos. Optimice los modelos para que sean buenos en ese trabajo. Y habrá enormes incentivos para hacerlo, con el fin de ayudarles a acelerar su investigación y la ventaja competitiva de sus laboratorios.

³⁴Por cierto, esto sugiere un punto importante en términos de la secuencia de los riesgos de la IA. Un modelo de amenaza común de IA que la gente señala es que los sistemas de IA desarrollan nuevas armas biológicas, y eso representa un riesgo catastrófico. Pero si la investigación de IA es más sencilla de automatizar que la investigación y el desarrollo de biología, podríamos tener una explosión de inteligencia antes de que tengamos amenazas biológicas extremas de IA. Esto es importante, por ejemplo, con respecto a si debemos esperar "disparos de advertencia biológica" a tiempo antes de que las cosas se vuelvan locas con la IA.

Seríamos capaces de ejecutar MILLONES de copias (y pronto a una velocidad humana de 10x+) de los investigadores automatizados de IA. Incluso para 2027, deberíamos esperar flotas de GPU en los 10 millones. Los clústeres de entrenamiento por sí solos deberían acercarse a ~ 3 OOM más grandes, lo que ya nos coloca en 10 millones + equivalentes a A100. Las flotas de inferencia deberían ser mucho más grandes aún. (Más sobre todo esto en [IIa. Corriendo hacia el grupo del billón de dólares](#)).

Eso nos permitiría ejecutar muchos millones de copias de nuestros investigadores de IA automatizados, tal vez 100 millones de equivalentes de investigadores humanos, funcionando día y noche. Hay algunas suposiciones que fluyen en los números exactos, incluyendo que los humanos "piensan" a 100 tokens/minuto (solo una estimación aproximada de orden de magnitud, por ejemplo, considere su monólogo interno) y la extrapolación de las tendencias históricas y las leyes de escalamiento de Chinchilla sobre los costos de inferencia por token para los modelos de frontera que permanecen en el mismo estadio.³⁵ También nos gustaría reservar algunas de las GPU para ejecutar experimentos y entrenar nuevos modelos. Cálculo completo en una nota a pie de página.³⁶

Otra forma de pensarlo es que, *dadas las flotas de inferencia en 2027, deberíamos ser capaces de generar todo un par de tokens en Internet, todos los días.*³⁷ En cualquier caso, las cifras exactas no más allá de una simple demostración de plausibilidad.

³⁵ Como se señaló anteriormente, la API GPT-4 cuesta menos hoy que GPT-3 cuando se lanzó, lo que sugiere que el

La tendencia de las ganancias de eficiencia de inferencia es lo suficientemente rápida como para mantener los costos de inferencia más o menos constantes, incluso cuando los modelos se vuelven mucho más poderosos. Del mismo modo, ha habido enormes ganancias en costos de inferencia solo en el año desde que se lanzó GPT-4; por ejemplo, la versión actual de Gemini 1.5 Pro supera al GPT-4 original, mientras que es 10 veces más barata.

También podemos fundamentar esto un poco más considerando las leyes de escalado de Chinchilla. En las leyes de escalamiento de Chinchilla, el tamaño del modelo y, por lo tanto, los costos de inferencia, crecen con la raíz cuadrada del costo de entrenamiento.

es decir, la mitad de los OOM del escalado OOM de proceso efectivo. Sin embargo, en

en el artículo anterior, sugerí que la eficiencia algorítmica avanzaba aproximadamente al mismo ritmo que el escalado de computación, es decir, representaba aproximadamente la mitad de la

las OOM del escalado vertical de proceso efectivo. Si estas victorias algorítmicas también se traducen en eficiencia de inferencia, eso significa que las eficiencias algorítmicas compensarían el aumento ingenuo en el costo de inferencia.

En la práctica, las eficiencias informáticas de entrenamiento a menudo, pero no siempre, se traducen en ganancias de eficiencia de inferencia. Sin embargo, también hay muchas victorias en la eficiencia de la inferencia que no son ganancias en la eficiencia del entrenamiento. Así que, al menos en términos aproximados, asumir que el \$/token de los modelos fronterizos se mantiene más o menos similar no parece una locura.

(Por supuesto, usarán más tokens, es decir, más cómputo en tiempo de prueba. Pero eso ya es parte del cálculo aquí, al fijar el precio de los equivalentes humanos en 100 tokens/minuto).

³⁶ GPT4Tson alrededor de \$ 0.03 / 1K tokens. Supusimos que tendríamos decenas de millones de equivalentes de A100, que costaban ~ \$ 1 hora por GPU si es equivalente a A100. Si usáramos los costos de la API para trans-GPU tardías en tokens generados, lo que implica 10 millones de GPU

* \$1/GPU-hora * 33K tokens/\$ = ~ un

billón de tokens/hora. Supongamos que un humano hace 100 tokens/min de pensamiento, eso significa que un equivalente humano es de 6.000 tokens/hora. Un billón de tokens/hora dividido por 6.000 tokens/hora-humano = ~200 millones de equivalentes humanos— es decir, como si dirigieran 200 millones de investigadores humanos, día y noche. (E incluso si reservamos la mitad de las GPU para el cálculo experimental, obtenemos 100 millones de equivalentes de investigadores humanos).

³⁷La nota a pie de página anterior estimaba ~1 billón de fichas/hora, es decir, 24 billones de fichas al día.

En el artículo anterior, señalé que un [CommonCrawl público deduplicado](#) tenía alrededor de 30 billones de tokens.

Además, es posible que nuestros investigadores de IA automatizada pronto puedan funcionar a una velocidad mucho más rápida que la humana:

- Al tomar algunas penalizaciones de inferencia, podemos compensar la ejecución de menos copias a cambio de ejecutarlas a una velocidad de serie más rápida. (Por ejemplo, podríamos pasar de ~5x velocidad humana a ~ 100 veces la velocidad humana al "solo" ejecutar 1 millón de copias de los investigadores automatizados.³⁸)

³⁸ Jacob Steinhardt estima que k^3 copias paralelas de un modelo pueden ser reemplazado por un solo modelo que es k^2

- Y lo que es más importante, la primera innovación algorítmica en la que trabajan los investigadores de IA está consiguiendo una aceleración de 10 o 100 veces. Gemini 1.5 Flash es ~10 veces más rápido que el GPT-4,39 lanzado originalmente solo un año después, al tiempo que proporciona un rendimiento similar al GPT-4 lanzado originalmente en los puntos de referencia de razonamiento. Si esa es la aceleración algorítmica que unos pocos investigadores humanos pueden encontrar en un año, los investigadores de IA automatizada podrán encontrar victorias similares muy rápidamente.

¿Podrían nuestros millones de investigadores de IA automatizados (que pronto trabajarán a una velocidad humana 10 o 100 veces) comprimir el progreso algorítmico que los investigadores humanos habrían encontrado en una década en un año?

Es decir: espere 100 millones de investigadores automatizados, cada uno trabajando a una velocidad 100 veces superior a la humana, no mucho después de que empezemos a ser capaces de automatizar la investigación en IA. Cada uno de ellos podrá hacer el trabajo de un año en unos pocos días. El aumento en el esfuerzo de investigación, en comparación con unos pocos cientos de investigadores humanos insignificantes en un laboratorio de IA líder en la actualidad, que trabajan a una velocidad insignificante 1 veces humana, será extraordinario.

Esto podría acelerar drásticamente las tendencias existentes de progreso algorítmico, comprimiendo una década de avances en un año.

No necesitamos postular nada totalmente novedoso para que la investigación automática de la IA acelere intensamente el progreso de la IA. Al revisar los números del artículo anterior, vimos que el progreso algorítmico ha sido un motor central del progreso del aprendizaje profundo en la última década; observamos una línea de tendencia de ~0,5 OOMs/año solo en las eficiencias algorítmicas, con grandes ganancias algorítmicas adicionales al descojear en la parte superior. (Creo que la importancia del progreso algorítmico ha sido subestimada por muchos, y apreciarlo adecuadamente es importante para apreciar la posibilidad de una explosión de inteligencia).

más rápido, dadas [algunas matemáticas sobre las compensaciones de inferencia](#) con un esquema de mosaico (que teóricamente funciona incluso para k de 100 o más). Supongamos que las velocidades iniciales ya eran ~ 5 veces la velocidad humana (basada, por ejemplo, en la velocidad de GPT-4 en el lanzamiento). Entonces, al tomar esta penalización de inferencia (con $k = \sim 5$), podríamos ejecutar ~ 1 millón de investigadores de IA automatizados a ~ 100 veces la velocidad humana.

³⁹ Esta [fuente](#) compara el rendimiento de Flash a $\sim 6x$ GPT-4 Turbo, y GPT-4 Turbo fue más rápido que el GPT-4 original. Es probable que la latencia también sea aproximadamente 10 veces más rápida.

¿en lugar de? Eso sería 5+ OOM en un año.

No TE IMAGINES 100 millones de pasantes de ingeniería de software junior aquí (¡los tendremos antes, en los próximos dos años!). Los verdaderos investigadores de IA automatizada son muy inteligentes y, además de su ventaja cuantitativa bruta, los investigadores de IA automatizada tendrán otras enormes ventajas sobre los investigadores humanos:

- Serán capaces de leer todos y cada uno de los artículos de ML que se hayan escrito, han sido capaces de pensar profundamente en cada uno de los experimentos anteriores que se han realizado en el laboratorio, aprender en paralelo de cada una de sus copias y acumular rápidamente el equivalente a milenios de experiencia. Serán capaces de desarrollar intuiciones mucho más profundas sobre el aprendizaje automático que cualquier humano.
- Podrán escribir fácilmente millones de líneas de código complejo, mantener todo el código base en contexto y pasar décadas humanas (o más) revisando y volviendo a verificar cada línea de código en busca de errores y optimizaciones. Serán magníficamente competentes en todas las partes del trabajo.
- No tendrá que entrenar individualmente a cada automatizado Investigador de IA (de hecho, capacitar e incorporar a 100 millones de nuevos humanos contratados sería difícil). En su lugar, puedes enseñar e incorporar uno de ellos, y luego hacer réplicas. (Y no tendrás que preocuparte por la politiquería, la aclimatación cultural, etc., y trabajarán con la máxima energía y concentración día y noche).
- Un gran número de investigadores de IA automatizados podrán compartir contexto (tal vez incluso acceder al espacio latente de los demás, etc.), lo que permitirá una colaboración y coordinación mucho más eficientes en comparación con los investigadores humanos.
- Y, por supuesto, por muy inteligentes que fueran nuestros investigadores iniciales de IA automatizada, pronto seríamos capaces de dar más saltos OOM, produciendo modelos aún más inteligentes, incluso más capaces en la investigación automatizada de IA.

Imagínese un Alec Radford automatizado, *imagínese 100 millones de*

tomó Alec Radfords.⁴⁰ Creo que casi todos los investigadores de increíblemente talentoso

OpenAI estaría de acuerdo en que si tuvieran 10 Alec Radfords, por no hablar de 100 o 1.000 o 1 millón corriendo a una velocidad humana de 10 o 100 veces, podrían resolver muy rápidamente muchos de sus problemas. Incluso con varios otros cuellos de botella (más en un momento), comprimir una década de progreso algorítmico en un año como Un resultado que parece muy plausible. (Una aceleración de 10 veces a partir de un millón de veces más esfuerzo de investigación, lo que parece conservador en todo caso).

Eso sería 5+ OOM allí mismo. 5 OOM de victorias algorítmicas serían una escala similar a la que produjo el salto de GPT-2 a GPT-4, un salto de capacidad de ~un niño en edad preescolar a ~un estudiante de secundaria inteligente. Imagínense un salto cualitativo *por encima de* AGI, *por encima de* Alec Radford.

Es sorprendentemente plausible que pasemos de AGI a superinteligencia muy rápidamente, tal vez en 1 año.

Posibles cuellos de botella

Si bien esta historia básica es sorprendentemente sólida, y está respaldada por un exhaustivo trabajo de modelado económico, existen algunos cuellos de botella reales y plausibles que probablemente ralentizarán una explosión de inteligencia de investigación automatizada de IA.

Daré un resumen aquí, y luego los discutiré con más detalle en las secciones opcionales a continuación para aquellos interesados:

- **Computación limitada:** la investigación de IA no solo requiere buenas ideas, pensamiento o matemáticas, sino que también realiza experimentos para obtener una señal empírica sobre sus ideas. Un millón de veces más esfuerzo de investigación a través del trabajo de investigación automatizado no significará un progreso un millón de veces más rápido, porque el proceso seguirá siendo limitado, y el cálculo limitado para los experimentos será el cuello de botella. Aún así, incluso si esto no será una aceleración de 1,000,000 de veces, me resulta difícil imaginar que los investigadores de IA automatizada no puedan usar el cómputo *al menos* 10 veces más efectivamente: podrán obtener una intuición increíble de ML (¡habiendo internalizado toda la literatura de ML y todos los experimentos anteriores en cada ejecución!) y

⁴⁰ Alec Radford es un jugador

y prolífico investigador/ingeniero en OpenAI, detrás de muchos de los avances más importantes, aunque algunos pasa desapercibido.

el equivalente a siglos de tiempo de reflexión para descubrir exactamente el experimento correcto para ejecutar, configurarlo de manera óptima y obtener el máximo valor de la información; Podrán pasar siglos equivalentes al tiempo de ingeniería antes de ejecutar incluso pequeños experimentos para evitar errores y acertar en el primer intento; Pueden hacer concesiones para economizar en computación centrándose en las mayores ganancias; y podrán probar toneladas de experimentos a menor escala (y dadas las escalas informáticas efectivas para entonces, "a menor escala" significa poder entrenar 100.000 modelos de nivel GPT-4 en un año para probar avances en arquitectura). Algunos investigadores e ingenieros humanos son capaces de producir 10 veces más progreso que otros, incluso con la misma cantidad de cómputo, y esto debería aplicarse aún más a los investigadores de IA automatizados. Creo que este es el cuello de botella más importante, y lo abordaré con más profundidad a continuación.

- **Complementariedades/cola larga:** Una lección clásica de la economía (cf. la enfermedad del crecimiento de Baumol) es que si puedes automatizar, digamos, el 70% de algo, obtienes algunas ganancias, pero rápidamente el 30% restante se convierte en tu cuello de botella. Para cualquier cosa que no llegue a la automatización total, por ejemplo, copilotos realmente buenos, los investigadores humanos de IA seguirían siendo un cuello de botella importante, lo que haría que el aumento general de la tasa de progreso algorítmico fuera relativamente pequeño. Además, es probable que se requiera una larga cola de capacidades para automatizar la investigación de IA: el último 10% del trabajo de un investigador de IA podría ser particularmente difícil de automatizar. Esto podría suavizar un poco el despegue, aunque mi mejor conjetura es que esto solo retrasa las cosas un par de años. Tal vez la velocidad de los modelos de 2026/27 sea la proto-automatizada investigadora, se necesitan uno o dos años más para que se descomponga un poco el cojeo final, un modelo algo mejor, se acelere la inferencia y se resuelvan los problemas para llegar a la automatización completa, y finalmente para 2028 obtendremos la aceleración 10x (y la superinteligencia para fines de la década).
- **Límites inherentes al progreso algorítmico:** ¿Quizás otros 5 OOM de eficiencia algorítmica serán fundamentalmente imposibles? Lo dudo. Si bien definitivamente habrá una parte superior
Límites⁴¹ si obtuvimos 5 OOM en la última década, deberíamos

⁴¹ 25 OOM de progreso algorítmico en

Probablemente esperen al menos otra década de progreso

top de GPT-4, por ejemplo, son claramente imposibles: eso implicaría que sería posible entrenar un modelo de nivel GPT-4 con solo un puñado de FLOP.

para que sea posible. Más directamente, las arquitecturas actuales y los algoritmos de entrenamiento son todavía muy rudimentarios, y parece que deberían ser posibles esquemas mucho más eficientes. Las clases de referencia biológicas también admiten algoritmos mucho más eficientes que son plausibles.

- ***Las ideas se vuelven más difíciles de encontrar, por lo que los investigadores de IA automatizada SIMPLEMENTE MANTENDRÁN, EN LUGAR DE ACELERAR, EL RITMO ACTUAL de progreso:*** una objeción es que, aunque la investigación automatizada aumentaría mucho el esfuerzo de investigación efectivo, las ideas también se vuelven más difíciles de encontrar. Es decir, mientras que hoy en día solo se necesitan unos pocos cientos de investigadores de primer nivel en un laboratorio para mantener 0,5 OOM/año, a medida que agotamos la fruta madura, se necesitará cada vez más esfuerzo para mantener ese progreso, por lo que los 100 millones de investigadores automatizados serán simplemente lo que se necesita para mantener el progreso. Creo que este modelo básico es correcto, pero los datos empíricos no cuadran: la magnitud del aumento en el esfuerzo de investigación —un millón de veces— es mucho, mucho mayor que las tendencias históricas del crecimiento en el esfuerzo de investigación que ha sido necesario para mantener el progreso. En términos de modelado económico, es una extraña "suposición al filo de la navaja" suponer que el aumento en el esfuerzo de investigación de la automatización será *suficiente* para mantener el progreso constante.
- ***Las ideas se vuelven más difíciles de encontrar y hay rendimientos decrecientes, por lo que la explosión de inteligencia se desvanecerá rápidamente:*** Relacionado con la objeción anterior, incluso si los investigadores de IA automatizada lideran A un estallido inicial de progreso, el hecho de que se pueda mantener un progreso rápido depende de la forma de la curva de rendimientos decrecientes del progreso algorítmico. Una vez más, mi mejor lectura de la evidencia empírica es que los exponentes se sacuden a favor del progreso explosivo/acelerado. En cualquier caso, el tamaño del impulso único (de 100 a 100 millones de investigadores de IA) probablemente supere los rendimientos decrecientes aquí para al menos un buen número de OOM de progreso algorítmico, aunque, por supuesto, no puede ser *indefinidamente* autosostenible.

En general, estos factores pueden ralentizar un poco las cosas: las versiones más extremas de la explosión de inteligencia (por ejemplo, de la noche a la mañana) parecen inverosímiles. Y pueden resultar en un tiempo algo más largo

(tal vez tengamos que esperar uno o dos años más de los investigadores más lentos y protoautomatizados a los verdaderos Alec Radfords autoacoplados, antes de que las cosas comiencen con toda su fuerza). Pero ciertamente no descartan una explosión de inteligencia muy rápida. Un año, o a lo sumo unos pocos años, pero tal vez incluso unos pocos meses, en el que pasemos de investigadores de IA totalmente automatizados a sistemas de IA enormemente sobrehumanos debería ser nuestra principal expectativa.

Si prefiere omitir las discusiones en profundidad sobre los diversos cuellos de botella a continuación, [haga clic aquí para pasar a la siguiente sección.](#)

Proceso limitado para experimentos (opcional, con más profundidad)

La función de producción para el progreso algorítmico incluye dos factores de producción complementarios: el esfuerzo de investigación y el cálculo experimental. Los millones de investigadores de IA automatizados no tendrán más computación para ejecutar sus experimentos que los investigadores humanos de IA; Tal vez se queden sentados esperando a que terminen sus trabajos.

Este es probablemente el cuello de botella más importante de la explosión de inteligencia. En última instancia, se trata de una cuestión cuantitativa: ¿hasta qué punto es un cuello de botella? A fin de cuentas, me resulta difícil creer que los 100 millones de Alec Radfords no pudieran aumentar el producto marginal de la computación experimental en al menos 10 veces (y, por lo tanto, aún aceleraran el ritmo del progreso en 10 veces):

- *Hay muchas cosas que se pueden hacer con cantidades más pequeñas de proceso.* La forma en que funciona la mayoría de las investigaciones de IA es que se prueban cosas a pequeña escala y luego se extrapolan a través de leyes de escalado. (Muchos avances históricos clave requirieron solo una cantidad muy pequeña de cómputo, p. ej. [el Transformer original](#) se entrenó con solo 8 GPU durante unos días). Y hay que tener en cuenta que con ~5 OOM de escalado de referencia en los próximos cuatro años, "pequeña escala" significará escala GPT-4: los investigadores de IA autoacoplados podrán ejecutar 100.000 experimentos de nivel GPT-4 en su grupo de entrenamiento en un año, y decenas de millones de

experimentos de nivel GPT-3. (Eso es mucho

de nuevas arquitecturas potencialmente revolucionarias que podrán probar!)

- Gran parte de la computación se destina a la validación a mayor escala de la ejecución final de preentrenamiento, asegurándose de que está obteniendo un grado de confianza lo suficientemente alto en las ganancias de eficiencia marginal para su producto principal anual, pero si está compitiendo a través de los OOM en la explosión de inteligencia, podría economizar y concentrarse en las ganancias realmente grandes.
- Como se discutió en el artículo anterior, a menudo se *obtienen enormes* ganancias al "destrabar" de los modelos de cómputo relativamente bajo. Estos no requieren grandes carreras previas al entrenamiento. Es muy plausible que esa explosión de inteligencia comience con la investigación automatizada de la IA, por ejemplo, descubriendo una forma de hacer RL en la parte superior que nos dé una OOM couple a través de victorias sin obstáculos (y luego nos vamos a las carreras).
- *A medida que los investigadores de IA automatizada encuentren eficiencias, eso les permitirá realizar más experimentos.* Recordemos la inferencia casi 1000 veces más barata en dos años para el rendimiento equivalente de MATH, y las ganancias de inferencia general 10 veces más baratas en el último año, [discutidas en el artículo anterior](#), a partir del mero progreso algorítmico humano. Lo primero que harán los investigadores de IA automatizada es encontrar rápidamente ganancias similares y, a su vez, eso les permitirá realizar 100 veces más experimentos sobre, por ejemplo, nuevos enfoques de RL. O podrán hacer rápidamente modelos más pequeños con un rendimiento similar en dominios relevantes (cf. discusión anterior de Gemini Flash, casi 100 veces más barato que GPT-4), lo que a su vez les permitirá realizar muchos más experimentos con estos modelos más pequeños (de nuevo, imag-
Es decir, que la mayoría de las personas que se encuentran en el campo de la información no tienen acceso a la información sobre la base de la información de la RL. Probablemente también haya otros voladizos, por ejemplo, los investigadores de IA automatizada

podrían desarrollar rápidamente esquemas de entrenamiento distribuidos mucho mejores para utilizar todas las GPU de inferencia (probablemente al menos 10 veces más cómputo allí mismo). De manera más general, cada OOM de ganancias de eficiencia de capacitación que encuentren les dará un OOM

Un proceso más eficaz para ejecutar experimentos.

- *Los investigadores de IA automatizados podrían ser mucho más eficientes.* Es difícil subestimar la cantidad de experimentos menos que tendrías que ejecutar si lo hicieras bien en el primer intento: sin errores retorcidos, siendo más selectivo sobre lo que estás ejecutando exactamente, etc. Imagínese que 1000 investigadores de IA automatizados pasan el equivalente a un mes revisando su código y obteniendo el experimento exacto justo antes de presionar ir. Le pregunté a algunos colegas del laboratorio de IA sobre esto y estuvieron de acuerdo: Podría ahorrar fácilmente de 3 a 10 veces de cómputo en la mayoría de los proyectos simplemente si pudiera evitar errores frívolos, hacer las cosas bien en el primer intento y ejecutar solo experimentos de alto valor de información.
- *Los investigadores de IA automatizada podrían tener intuiciones mucho mejores .*
 - Recientemente, estaba hablando con un pasante en un laboratorio de frontera; Dijeron que su experiencia dominante en los últimos meses era sugerir muchos experimentos que querían realizar, y su supervisor (un investigador principal) decía que ya podían predecir el resultado de antemano, por lo que no había necesidad. Los años de experimentos aleatorios de la investigadora principal jugando con modelos habían perfeccionado sus intuiciones sobre qué ideas funcionarían o no. Del mismo modo, parece que nuestros sistemas de IA podrían obtener fácilmente intuiciones sobrehumanas sobre los experimentos de aprendizaje automático: habrán leído toda la literatura de aprendizaje automático, podrán aprender de todos los demás resultados de los experimentos y pensar profundamente en ellos, podrían entrenarse fácilmente para predecir el resultado de millones de experimentos de aprendizaje automático, etc. Y tal vez una de las primeras cosas que hacen es construir una ciencia básica sólida de "predecir si esta gran escala El experimento tendrá éxito justo después de ver el primer 1% del entrenamiento, o justo después de ver la versión a menor escala de este experimento", y así sucesivamente.

- Por otra parte, más allá de las buenas intuiciones sobre la re-

direcciones de búsqueda, como ha señalado Jason Wei, hay increíbles retornos a tener grandes intuiciones sobre las docenas de hiperparámetros y detalles de un experimento. Jason llama a esta capacidad de hacer las cosas bien en el primer intento basado en la intuición "yolo runs". (Jason dice: "Lo que sí sé es que las personas que pueden hacer esto son seguramente 10-100 veces investigadores de IA").

Los cuellos de botella en la computación significarán que un millón de veces más investigadores no se traducirán en una investigación un millón de veces más rápida, por lo que no será una explosión de inteligencia de la noche a la mañana. Pero los investigadores de IA automatizada tendrán ventajas extraordinarias sobre los investigadores humanos, por lo que parece difícil imaginar que no puedan encontrar también una manera de usar el cómputo al menos 10 veces más eficiente/efectivo, y así 10 veces el ritmo del progreso algorítmico parece eminentemente plausible.

Me DETENDRÉ AQUÍ para reconocer quizás la formulación más convincente del contraargumento que he escuchado, de mi amigo James Bradbury: *si un mayor esfuerzo de investigación de ML aceleraría tan drásticamente el progreso, ¿por qué la actual comunidad académica de investigación de ML, que cuenta al menos con decenas de miles, no contribuye más al progreso de los laboratorios de frontera?* (Actualmente, parece que los equipos internos de los laboratorios, de quizás mil en total en todos los laboratorios, asumen la mayor parte de la carga del progreso algorítmico de frontera). Su argumento es que la causa es que el progreso algorítmico está plagado de cuellos de botella en la computación: los académicos simplemente no tienen suficiente computación.

Algunas respuestas:

- Ajustado por calidad, creo que los académicos son probablemente más miles, no decenas de miles (por ejemplo, sólo en las mejores universidades). Es probable que esto no sea sustancialmente más que los laboratorios combinados. (Y es *mucho* menos que los cientos de millones de investigadores que obtendríamos de la investigación automatizada de IA).

- Los académicos trabajan en las cosas equivocadas. Hasta muy

Recientemente (¿y quizás todavía hoy?), la gran mayoría de la comunidad académica de ML ni siquiera estaba trabajando en grandes modelos de lenguaje. En términos de académicos fuertes en el mundo académico que trabajan en grandes modelos lingüísticos, ¿podría ser significativamente menor que los investigadores de los laboratorios combinados?

- Incluso cuando los académicos trabajan en cosas como la formación previa al LLM, simplemente no tienen acceso al estado del arte: el gran cuerpo de conocimiento acumulado de toneladas de detalles sobre la formación de modelos de frontera dentro de los laboratorios. No saben qué problemas son realmente relevantes, o solo pueden aportar resultados puntuales con los que nadie puede hacer nada porque sus líneas de base estaban mal ajustadas (por lo que nadie sabe si lo suyo es realmente una mejora).
- Los académicos son mucho peores que los investigadores automatizados de IA: no pueden trabajar a una velocidad humana 10 o 100 veces mayor, no pueden leer e interiorizar todos los artículos de ML que se hayan escrito, no pueden pasar una década revisando cada línea de código, suplicarse a sí mismos para evitar cuellos de botella en la incorporación, etc.

Otro ejemplo compensatorio al argumento de los académicos: se rumorea que GDM tiene mucha más computación experimental que OpenAI y, sin embargo, no parece que GDM esté superando masivamente a OpenAI en términos de progreso algorítmico.

En general, espero que los investigadores automatizados tengan un estilo de investigación diferente que aproveche sus fortalezas y tenga como objetivo mitigar el cuello de botella de cómputo. *Creo que es razonable no estar seguro de cómo se desarrolla esto, pero no es razonable confiar en que no será factible para los modelos sortear el cuello de botella de cómputo solo porque sería difícil para los humanos hacerlo.*

- Por ejemplo, podrían dedicar mucho esfuerzo desde el principio a la construcción de una ciencia básica de "cómo predecir resultados a gran escala a partir de experimentos a menor escala". Y supongo que hay muchas cosas que podrían hacer que los humanos no

pueden hacer, por ejemplo, tal vez cosas más como
"predecir si este experimento a gran escala tendrá éxito
justo después de ver el

primer 1% de formación". Esto parece bastante factible si eres un investigador automatizado súper fuerte con intuiciones muy superhumanas y esto puede ahorrarte una tonelada de cómputo.

- Cuando imagino que los sistemas de IA automatizan la investigación de la IA, los veo como un cuello de botella en la computación, pero lo compensan en gran parte pensando, por ejemplo, 1000 veces más (y más rápido) de lo que lo harían los humanos, y pensando en un nivel de calidad más alto que los humanos (por ejemplo, debido a lo sobrehumano Intuiciones de ML de ser entrenado para predecir el resultado de millones de experimentos). A menos que sean mucho peores pensando que la ingeniería, creo que esto puede compensar mucho, y esto sería cualitativamente diferente de lo académico.

(Además de la computación de experimentos, existe el cuello de botella adicional de tener que ejecutar una gran ejecución de entrenamiento, algo que actualmente lleva meses. Pero probablemente se pueda economizar en ellos, haciendo solo un puñado durante el año de la explosión de inteligencia, dando mayores saltos de OOM para cada uno de los que hacen actualmente los laboratorios.⁴² O podrías "gastar" 1 de cada 5 OOM de eficiencia informática gana para realizar una ejecución de entrenamiento en días en lugar de meses).

IA para todos los dominios aún no se hayan resuelto, que la sociedad no se haya resuelto

Complementariedades y colas largas para la automatización al 100% (opcional, con más profundidad)

La objeción clásica de los economistas a que la automatización de la IA acelera el crecimiento económico es que las diferentes tareas son complementarias: así, por ejemplo, la automatización del 80% del trabajo que hacían los humanos en 1800 no condujo a una explosión de crecimiento o desempleo masivo, pero el 20% restante se convirtió en lo que todos los humanos hacían y siguió siendo el cuello de botella. (Véase, por ejemplo, un modelo de [este aquí](#)).

Creo que el modelo de los economistas es correcto. Pero un punto clave es que solo estoy hablando de una pequeña parte de la economía en la actualidad, en lugar de la economía en su conjunto. Es muy posible que la gente siga cortándose el pelo con normalidad durante este tiempo: es posible que la robótica aún no se haya desarrollado, que las

⁴² Tenga en cuenta que, si bien creo que esto es probable, es un poco aterrador: significa que en lugar de una serie bastante continua de grandes modelos, cada uno algo mejor que la generación anterior, la inteligencia del modelo posterior podría ser más discreta/discontinua. Es posible que solo hagamos una o un par de grandes carreras durante la explosión de inteligencia, acumulando múltiples OOM de avances algorítmicos encontrados a menor escala para cada uno.

Es posible que aún no se haya resuelto, etc., pero *lo harán* ser capaz de hacer investigación de IA. Como se discutió en el artículo anterior, creo que el curso actual del progreso de la IA nos está llevando esencialmente a trabajadores remotos tan inteligentes como los humanos más inteligentes; como se discute en este artículo, el trabajo de un investigador de IA parece estar totalmente dentro del alcance de lo que podría automatizarse por completo.

Aún así, en la práctica, espero una cola algo larga para llegar a una verdadera automatización del 100%, incluso para el trabajo de una IA re-
Por ejemplo, es posible que primero obtengamos sistemas que funcionen *casi* como un reemplazo de ingeniero, pero que aún necesiten cierta cantidad de supervisión humana.

En particular, espero que el nivel de las capacidades de IA sea algo desigual y pico en todos los dominios: podría ser un mejor programador que los mejores ingenieros sin dejar de tener puntos ciegos en algún subconjunto de tareas o habilidades; En el momento en que esté a nivel humano en lo que sea peor, ya será sustancialmente sobrehumano en dominios más fáciles de entrenar, como la codificación. (Esta es parte de la razón por la que creo que podrán usar el cómputo de manera más efectiva que los investigadores humanos. En el momento en que comience la automatización al 100% o la explosión de la inteligencia, ya tendrán enormes ventajas sobre los humanos en algunos dominios. Esto también tendrá implicaciones importantes para la superalineación en el futuro, ya que significa que tendremos que alinear sistemas que son significativamente sobrehumanos en muchos dominios con el fin de alinear incluso a los primeros investigadores de IA automatizada).

Pero no esperaría que esa fase durara más de unos pocos años; dado el ritmo del progreso de la IA, creo que probablemente solo sería una cuestión de "destrabar" adicional (eliminar alguna limitación obvia de los modelos que le impedía hacer la última milla) u otra generación de modelos para llegar hasta el final.

En general, esto podría suavizar un poco el despegue. En lugar de 2027 AGI → 2028 Superintelligence, podría parecerse más a:

- 2026/27: Proto-ingeniero automatizado, pero puntos ciegos en

otras áreas. Acelera el trabajo entre 1,5 y 2 veces; El progreso comienza a acelerarse gradualmente.

- 2027/28: Los investigadores protoautomatizados pueden automatizar >90%. Quedan algunos cuellos de botella humanos y contratiempos en la coordinación de una organización gigante de investigadores automatizados que deben resolverse, pero esto ya acelera el progreso en 3 veces +. De este modo, el resto de la necesidad de "destrabar" nos lleva rápidamente al resto del camino hacia la automatización al 100%.
- 2028/29: 10x+ ritmo de progreso → superinteligencia.

Eso sigue siendo muy rápido. . .

Límites fundamentales del progreso algorítmico (opcional, con más profundidad)

Probablemente haya un límite real en la cantidad de progreso algorítmico que es físicamente posible. (Por ejemplo, 25 OOM de progreso del algoritmo parece imposible, ya que eso implica ser capaz de entrenar un sistema de nivel GPT-4 en menos de ~10 FLOPS.⁴³) Pero algo así como 5 OOM parece muy resultados que

mucho en el reino de las posibilidades; Una vez más, eso solo requeriría otra década de eficiencias algorítmicas de tendencia (sin contar las ganancias algorítmicas de la liberación).

⁴³ A pesar de que podrías obtener

con pequeños ajustes, mientras que hay docenas de áreas en las que se puede ser mucho más eficiente.

Intuitivamente, no parece que hayamos agotado aún todos los frutos que están al alcance de la mano, dado lo simples que son los mayores avances, y lo rudimentarias y obviamente obstaculizadas que parecen ser las arquitecturas y técnicas de entrenamiento actuales. Por ejemplo, creo que es bastante plausible que nos abramos camino hacia la AGI a través de los sistemas de IA. A través de la cadena de pensamiento, la mayoría de las personas que "piensan en voz alta" no pueden ser capaces de pensar en voz alta. Pero *seguramente* esta no es la forma más eficiente de hacerlo, *seguramente* algo que haga este razonamiento a través de estados internos/recurrencia/etc. sería mucho más eficiente. O considere la computación adaptativa: Llama 3 todavía gasta tanto cómputo en predecir el token "y" como en la respuesta a alguna pregunta complicada, que parece claramente subóptima. Estamos obteniendo enormes ganancias algorítmicas de OOM incluso

¡Se necesitarían 25 OOM más de hardware con la arquitectura actual!

Es probable que se puedan encontrar arquitecturas y procedimientos de entrenamiento eficientes.

Las referencias biológicas también sugieren un gran margen de maniobra. El rango de inteligencia humana es muy amplio, por ejemplo, con solo pequeños ajustes en la arquitectura. Los seres humanos tienen un número similar de neuronas que los demás animales, a pesar de que los humanos son mucho más inteligentes que esos animales. Y los modelos actuales de IA siguen estando muy lejos de la **eficiencia de un cerebro humano**; pueden aprender con una pequeña fracción de los datos (y, por lo tanto, una pequeña fracción de "cómputo") que los modelos de IA, lo que sugiere un enorme margen para nuestros algoritmos y arquitectura.

Las ideas se vuelven más difíciles de encontrar y los rendimientos decrecientes (opcional, con más profundidad)

A medida que recoges la fruta que está al alcance de la mano, las ideas se vuelven más difíciles de encontrar. Esto es cierto en cualquier **dominio del progreso tecnológico**. Esencialmente, vemos una línea recta en una curva log-log: $\log(\text{progreso})$ es una función de $\log(\text{esfuerzo acumulado de investigación})$. Cada OOM

de un mayor progreso requiere un mayor esfuerzo de investigación que el último OOM.

Esto lleva a dos objeciones a la explosión de inteligencia:

1. La investigación automatizada de IA será simplemente lo necesario para mantener el progreso (en lugar de acelerarlo drásticamente).
2. Una explosión de inteligencia puramente algorítmica no se sostendría / se desvanecería rápidamente a medida que el progreso algorítmico se vuelve más difícil de encontrar / se obtienen rendimientos marginales decrecientes.

Pasé mucho tiempo pensando en este tipo de modelos en una vida pasada, cuando estaba investigando en economía. (En particular, La **teoría del crecimiento semiendógeno** es el modelo estándar del progreso tecnológico, que capta estas dos dinámicas que compiten entre sí entre el esfuerzo creciente de investigación y las ideas que se vuelven más difíciles de encontrar).

En resumen, creo que el modelo subyacente detrás de estas objeciones es sólido, pero cómo se desarrolla es una cuestión empírica, y creo que se equivocan en los empíricos.

La pregunta clave es, esencialmente: por cada 10 veces de progreso, ¿el progreso adicional se vuelve *más* o *menos* de 10 veces más difícil? La matemática de la servilleta (en la línea de cómo se hace esto en la [literatura económica](#)) nos ayuda a encuadernar esto.

- Supongamos que nos tomamos en serio la tasa de tendencia de $\sim 0,5$ OOM/año del progreso algorítmico; Eso implica un avance de 100 veces en 4 años.
- Sin embargo, el número de empleados / esfuerzo de investigación ajustado por calidad en un laboratorio de IA líder determinado definitivamente ha crecido < 100 veces en 4 años. *Tal vez* se haya incrementado 10 veces (de 10 a 100 personas que trabajan en cosas relevantes en un laboratorio determinado), pero incluso eso no está claro ajustado a la calidad.
- Y, sin embargo, el progreso algorítmico parece sostenerse.

Así, en respuesta a la objeción 1, podemos señalar que la \sim Un aumento de \sim millones de veces en el esfuerzo de investigación será simplemente un aumento mucho mayor de lo que sería meramente necesario para sostener el progreso. Tal vez necesitaríamos del orden de miles de investigadores trabajando en investigaciones relevantes en un laboratorio en 4 años para mantener el progreso; los 100 millones de Alec Radfords seguirían siendo un *aumento enorme*, y seguramente conducen a una aceleración masiva. Es solo una extraña suposición "al filo de la navaja" pensar que la investigación automatizada sería *suficiente* para mantener el ritmo de progreso existente. (Y eso sin contar el pensamiento a 10 veces la velocidad humana y todas las otras ventajas que tendrán los sistemas de IA sobre los investigadores humanos).

En respuesta a la objeción 2, podemos señalar dos cosas:

- En primer lugar, la condición matemática señalada anteriormente. Dado que, según nuestras

matemáticas de servilleta, el esfuerzo de investigación ajustado por calidad necesitaba crecer «100 veces mientras que nosotros lo hacíamos 100 veces del progreso algorítmico, parece que la

la forma de la curva de retorno se sacude a favor de la Mantener el progreso.⁴⁴

- En segundo lugar, la curva de retornos ni siquiera *necesita* sacudirse a favor de una reacción en cadena totalmente sostenida para que obtengamos un aumento limitado pero de muchos OOM de progreso algorítmico. Esencialmente, no tiene por qué ser un "efecto de crecimiento"; Un "efecto de nivel" lo suficientemente grande sería suficiente. Es decir, un aumento de un millón de veces en el esfuerzo de investigación (combinado con las muchas otras ventajas que tendrían los investigadores automatizados sobre los investigadores humanos) ser un impulso único tan grande que, incluso si la reacción en cadena no es totalmente autosuficiente, podría conducir a una considerable (muchos OOM) ganancia única.⁴⁵

En la red, aunque obviamente no será ilimitado y tengo mucha incertidumbre sobre hasta dónde llegará, creo que algo como una explosión de inteligencia de 5 OOM puramente por ganancias algorítmicas / investigación automatizada de IA parece muy plausible.

Tom Davidson y Carl Shulman también han analizado los aspectos empíricos de esto en un marco de modelado de crecimiento y han llegado a conclusiones similares. *Epoch AI también ha realizado algunos trabajos recientes* sobre los datos empíricos, llegando también a la conclusión de que los retornos empíricos de la investigación y el desarrollo algorítmicos favorecen el crecimiento explosivo, con una *útil descripción de las implicaciones*.

El poder de la superinteligencia

Independientemente de que se esté de acuerdo o no con la forma más fuerte de estos argumentos, ya sea que tengamos una explosión de inteligencia de < y un año, o que tome unos pocos años, está claro: debemos confrontar la posibilidad de la superinteligencia.

Los sistemas de IA que probablemente tendremos *a finales de*

⁴⁴ Progreso 100x → 100 veces más esfuerzo de investigación autoacoplado, pero tú, digamos, solo requirió 10 veces más esfuerzo de investigación para mantener Va y hace las siguientes 100 veces, por lo que los rendimientos son lo suficientemente buenos como para mantener un progreso explosivo.

⁴⁵ De manera análoga, en semiendógenos teoría del crecimiento económico, impulsando inversión en ciencia del 1% al 20% de la *esta década* serán inimaginablemente poderosos.

- Por supuesto, serán *cuantitativamente* sobrehumanos. En nuestras flotas

El PIB no hará que las tasas de crecimiento sean más altas para siempre —con el tiempo, los rendimientos decrecientes llevarían a que las cosas volvieran a la antigua tasa de crecimiento—, pero el "efecto nivel" al que conduciría sería tan grande como para acelerar drásticamente el crecimiento durante décadas.

de cientos de millones de GPU para finales de la década, seremos capaces de dirigir una civilización de miles de millones de ellas, y serán capaces de "pensar" órdenes de magnitud más rápido que los humanos. Serán capaces de dominar rápidamente cualquier dominio, escribir billones de líneas de código, leer todos los trabajos de investigación en todos los campos científicos que se hayan escrito (¡serán perfectamente interdisciplinarios!) y escribir otros nuevos antes de que hayas pasado del resumen de

Uno, aprender de la experiencia paralela de cada una de sus copias, ganar miles de millones de años de experiencia equivalente a humanos con alguna nueva innovación en cuestión de semanas, trabajar el 100% del tiempo con la máxima energía y concentración y no ser frenado por ese compañero de equipo que se está quedando atrás, y así sucesivamente.

- Y lo que es más importante, pero más difícil de imaginar, serán *cualitativamente* sobrehumanos. Como un ejemplo limitado de esto, las carreras de RL a gran escala han sido capaces de producir comportamientos completamente novedosos y creativos más allá de la comprensión humana, como la [famosa jugada 37 en AlphaGo vs. Lee Sedol](#). La superinteligencia será así en muchos dominios. Encontrará exploits en el código humano demasiado sutiles para que cualquier humano los note, y generará un código demasiado complicado para que cualquier humano lo entienda, incluso si el modelo pasó décadas tratando de explicarlo. Los problemas científicos y tecnológicos extremadamente difíciles en los que un ser humano estaría atrapado durante décadas les parecerán *tan obvios*. Seremos como estudiantes de secundaria atrapados en la física newtoniana mientras exploran la mecánica cuántica.

Como ejemplo de lo *salvaje* que podría ser esto, mira algunos videos de Youtube de speedruns de videojuegos, como este [devencer a Minecraft en 20 segundos](#). (Si no tienes idea de lo que está pasando

En este video, estás en buena compañía; incluso la mayoría de los jugadores normales de Minecraft casi no tienen ni idea de lo que está pasando).

Ahora imagina que esto se aplica a todos los dominios de la ciencia, la tecnología y la economía. Las barras de error aquí, por supuesto, son extremadamente grandes. Aun así, *esto está ocurriendo*, y es importante considerar qué tan importante sería esto.

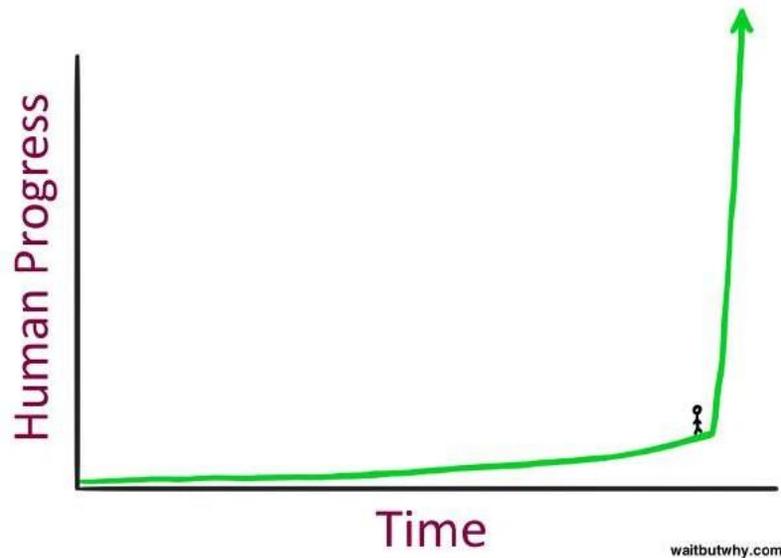


Figura 25: ¿Qué se siente al estar aquí?
Ilustración de Wait But Why/Tim Urban.

En la explosión de la inteligencia, el progreso explosivo fue inicialmente solo en el estrecho dominio de la investigación automatizada de IA. A medida que adquirimos superinteligencia y aplicamos nuestros miles de millones de agentes (ahora superinteligentes) a la investigación y el desarrollo en muchos campos, espero que el progreso explosivo se amplíe:

- **Una explosión de capacidades de IA.** Tal vez nuestros AGI iniciales tenían limitaciones que les impedían automatizar completamente el trabajo en algunos otros dominios (en lugar de solo en el dominio de la investigación de IA); la investigación automatizada de IA resolverá rápidamente esto, permitiendo la automatización de todo el trabajo cognitivo.
- **Resuelve la robótica.** La superinteligencia no permanecerá puramente cognitiva por mucho tiempo. Lograr que la robótica funcione bien es principalmente un problema de algoritmos de ML (en lugar de un problema de hardware), y es probable que nuestros investigadores de IA automatizada puedan resolverlo (¡más a continuación!). Las fábricas pasarían de ser administradas por humanos, a ser dirigidas por IA utilizando trabajo físico humano, y pronto serían completamente administradas por enjambres de robots.
- **Acelerar drásticamente el progreso científico y tecnológico.** Sí, Einstein por sí solo no pudo desarrollar la neurociencia y construir una industria de semiconductores, pero mil millones de científicos, ingenieros, tecnólogos y robots

automatizados superinteligentes

técnicos (con los robots moviéndose a 10 veces o más personas ¡Velocidad!)⁴⁶ lograría avances extraordinarios en muchos campos en el espacio de años. (Aquí hay una [buena historia corta](#) que visualiza cómo podría ser la investigación y el desarrollo impulsados por la IA). Los mil millones de superinteligencias serían capaces de comprimir en años el esfuerzo de investigación y desarrollo que los investigadores humanos habrían realizado en el próximo siglo. Imagínese si el progreso tecnológico del siglo XX se comprimiera en menos de una década. Habríamos pasado de volar siendo considerado un espejismo, a los aviones, a un hombre en la luna y a los misiles balísticos intercontinentales en cuestión de años. *Así es como espero que sea la década de 2030 en la ciencia y la tecnología.*

⁴⁶ Los robots de velocidad 10x haciendo La investigación y el desarrollo en el mundo real es la "versión lenta"; en realidad, las superinteligencias tratarán de hacer tanta investigación y desarrollo como sea posible en simulación, como [AlphaFold](#) la fabricación de "gemelos digitales".

- **Una explosión industrial y económica.** Un progreso tecnológico extremadamente acelerado, combinado con la capacidad de automatizar todo el trabajo humano, podría acelerar drásticamente el crecimiento económico (pensemos: fábricas de robots autorreplicantes rápidamente cubriendo todo el desierto de Nevada⁴⁷). El aumento del crecimiento⁴⁷ ¿Por qué no es "factorio-world"— construir Probablemente no sería sólo del 2% anual al 2,5% anual, sino que se trataría de un cambio fundamental en el régimen de crecimiento, más comparable al cambio histórico de un crecimiento muy lento a un par de puntos porcentuales anuales con la revolución industrial. Podríamos ver tasas de crecimiento económico del 30% anual y más, muy posiblemente múltiples duplicaciones al año. Esto se deduce claramente de los [modelos de crecimiento económico](#) de los economistas. Sin duda, esto puede retrasarse por las fricciones sociales; la regulación arcana podría asegurar que los abogados y los médicos sigan necesitando ser humanos, incluso si los sistemas de IA fueran mucho mejores en esos trabajos; seguramente se arrojará arena a los engranajes de las fábricas robóticas en rápida expansión a medida que la sociedad se resiste al ritmo del cambio; y tal vez queramos hacerlo retener niñeras humanas; todo lo cual frenaría el crecimiento de las estadísticas generales del PIB. Aun así, en cualquier ámbito en el que eliminemos las barreras creadas por el hombre (por ejemplo, la competencia podría obligarnos a hacerlo para la producción militar), veríamos una explosión industrial.

Una fábrica, que produce más fábricas, que produce aún más fábricas, que duplica las fábricas hasta que finalmente todo su planeta se cubre rápidamente de fábricas, ¿es posible hoy en día? Bueno, la mano de obra está limitada: se puede acumular capital (fábricas, herramientas, etc.), pero eso se topa con rendimientos decrecientes porque está limitado por una fuerza de trabajo fija. Dado que los robots y los sistemas de IA pueden automatizar completamente el trabajo, eso elimina esa restricción; Las fábricas robóticas podrían producir más fábricas robóticas en un ~sin restricciones, lo que provoca una explosión industrial. Vea más modelos de crecimiento económico de esto [aquí](#).

Modo de crecimiento	El dátil comenzó a dominar	Tiempo de duplicación de la economía mundial (años)
Caza	2.000.000 a.C.	230,000
Agricultura	4700 a.C.	860
Ciencia/comercio	Año 1730 d.C.	58
Industria	Año 1903 d.C.	15
¿Superinteligencia?	¿Año 2030 d.C.?	???

tenían un poder divino, pero la ventaja tecnológica del Viejo Mundo y una ventaja en la estrategia.

- **Proporcionar una ventaja militar decisiva y abrumadora.**

Incluso la superinteligencia cognitiva temprana podría ser suficiente aquí; tal vez algún esquema de piratería sobrehumana pueda desactivar a los ejércitos adversarios. En cualquier caso, el poder militar y el progreso tecnológico han estado estrechamente ligados históricamente, y con el progreso tecnológico extraordinariamente rápido vendrán las revoluciones militares concomitantes. Los enjambres de drones y los roboejércitos serán un gran problema, pero son solo el comienzo; deberíamos esperar tipos de armas completamente nuevos, desde nuevas armas de destrucción masiva hasta invulnerables defensas antimisiles basadas en láser y cosas que aún no podemos comprender. En comparación con los arsenales anteriores a la superinteligencia, será como los ejércitos del siglo XXI.

A pesar de todo, la mayoría de las personas que luchan contra una brigada de caballos y bayonetas del siglo XIX. (Discuto cómo la superinteligencia podría conducir a una ventaja militar decisiva en un artículo posterior).

- **Ser capaz de derrocar al gobierno de los Estados Unidos.**

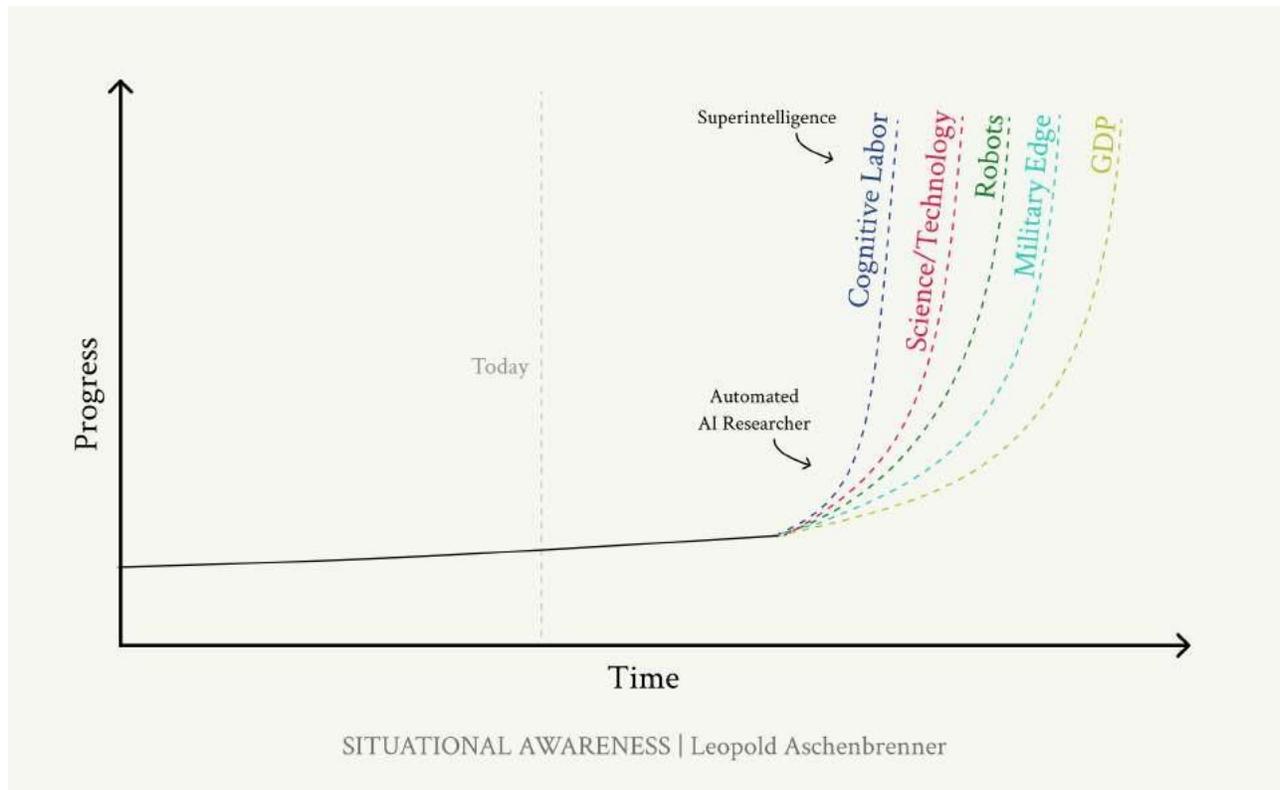
Quienquiera que controle la superinteligencia muy posiblemente tendrá suficiente poder para arrebatar el control a las fuerzas anteriores a la superinteligencia. Incluso sin robots, la pequeña civilización de las superinteligencias sería capaz de hackear cualquier sistema militar, electoral, televisivo, etc. indefenso, persuadir astutamente a generales y electorados, superar económicamente a los estados-nación, diseñar nuevas armas biológicas sintéticas y luego pagar a un humano en bitcoin para sintetizar.

tamaño, y así sucesivamente. **A principios de 1500**, Cortés y unos 500 españoles conquistaron el imperio azteca de varios millones de habitantes; Pizarro y ~300 españoles conquistaron el imperio inca de varios millones; Alfonso y ~1000 portugueses conquistaron el Océano Índico. No

Tabla 3: Un cambio en el régimen de crecimiento no es inédito: a medida que la civilización pasó de la caza a la agricultura, al florecimiento de la ciencia y el comercio, para la industria, el ritmo del crecimiento económico mundial se aceleró. La superinteligencia podría iniciar otro cambio en el modo de crecimiento. Basado en Robin Hanson

"Crecimiento a largo plazo como una secuencia demodos exponenciales".

La astucia gic y diplomática condujo a una ventaja absolutamente decisiva. La superinteligencia podría tener un aspecto similar.



Robots. Una objeción común a afirmaciones como las que se presentan aquí es que, incluso si la IA puede realizar tareas cognitivas, la robótica se está quedando muy atrás y, por lo tanto, será un freno para cualquier impacto en el mundo real.

Solía simpatizar con esto, pero me he convencido de que los robots no serán una barrera. Durante años, la gente afirmó que los robots eran un problema de hardware, pero el hardware de los robots está en camino de resolverse.

Cada vez está más claro que los robots son un problema de los algoritmos de aprendizaje automático. Los LLM tenían una forma mucho más fácil de arrancar: tenías todo un Internet para preentrenarte. No existe un conjunto de datos igualmente grande para las acciones de los robots, por lo que se requieren enfoques más ingeniosos (por ejemplo, el uso de modelos multimodales como base, y luego el uso de datos sintéticos/simulación/RL inteligente) para entrenarlos.

Figura 26: El crecimiento explosivo comienza en el dominio más estrecho de la investigación y el desarrollo de IA; a medida que apliquemos la superinteligencia a la investigación y el desarrollo en otros campos, el crecimiento explosivo se ampliará.

Hay una tonelada de energía dirigida a resolver esto ahora. Pero incluso si no lo resolvemos antes de AGI, nuestros cientos de millones de AGIs/superinteligencias serán *increíbles* investigadores de IA (¡como es el argumento central de este artículo!), y parece muy probable que descubran el ML para hacer que los robots asombrosos funcionen.

Como tal, si bien es plausible que los robots puedan causar algunos años de retraso (resolver los problemas de ML, realizar pruebas en el mundo físico de una manera que sea fundamentalmente más lenta que las pruebas en simulación, aumentar la producción inicial de robots antes de que los robots puedan construir fábricas por sí mismos, etc.) —No creo que sea más que eso.

Es difícil predecir cómo se desarrollará TODO ESTO durante la década de 2030 (y es una historia para otro momento). Pero una cosa, al menos, está clara: nos sumergiremos rápidamente en la situación más extrema a la que se haya enfrentado la humanidad.

Los sistemas de IA a nivel humano, AGI, serían muy importantes por derecho propio, pero en cierto sentido, simplemente serían una versión más eficiente de lo que ya sabemos. Pero, muy Plausiblemente, en sólo un año, haríamos la transición a sistemas mucho más alienígenas, sistemas cuya comprensión y habilidades, cuyo poder bruto, superaría incluso a los de la humanidad combinados. Existe una posibilidad real de que perdamos el control, ya que nos vemos obligados a ceder la confianza a los sistemas de IA durante esta rápida transición.

En términos más generales, todo comenzará a suceder *increíblemente rápido*. Y el mundo empezará a volverse loco. Supongamos que hubiéramos pasado por los puntos álgidos geopolíticos y los peligros provocados por el hombre del siglo XX en cuestión de años; Ese es el tipo de situación que deberíamos esperar después de la superinteligencia. Al final, los sistemas de IA superinteligentes estarán dirigiendo nuestras fuerzas armadas y economía. Durante toda esta locura, tendríamos muy poco tiempo para tomar las decisiones correctas. Los desafíos serán inmensos. Se necesitará todo lo que tenemos para salir adelante de una pieza.

La explosión de inteligencia y el período inmediatamente posterior a la superinteligencia serán uno de los períodos más volátiles, tensos, peligrosos y salvajes de la historia de la humanidad.

Y para el final de la década, es probable que estemos en medio de ella.

Enfrentarse a LA posibilidad de una explosión de inteligencia —el surgimiento de la superinteligencia— a menudo se hace eco de los primeros debates en torno a la posibilidad de una reacción nuclear en cadena.

y la bomba atómica que permitiría. H.G. Wells predijo la bomba atómica en una novela de 1914. Cuando Szilard concibió por primera vez la idea de una reacción en cadena en 1933, no pudo convencer a nadie de ella; Era pura teoría. Una vez que la fisión fue descubierta empíricamente en 1938, Szilard se asustó de nuevo y abogó enérgicamente por el secreto, y algunas personas comenzaron a despertar ante la posibilidad de una bomba. Einstein no había considerado la posibilidad

de una reacción en cadena, pero cuando Szilard se enfrentó a él, se apresuró a ver las implicaciones y estuvo dispuesto a hacer cualquier cosa que fuera necesario; Estaba dispuesto a hacer sonar la alarma y no tenía miedo de parecer tonto. Pero Fermi, Bohr y la mayoría de los científicos pensaron que lo "conservador" era minimizarlo, en lugar de tomar en serio las extraordinarias implicaciones de la posibilidad de una bomba. El secretismo (para evitar compartir sus avances con los alemanes) y otros esfuerzos sin cuartel parecían absurdos a ellos. Una reacción en cadena sonaba demasiado loca. (Incluso cuando, como se vio, una bomba estaba a solo media década de convertirse en realidad).

Debemos enfrentarnos una vez más a la posibilidad de una reacción en cadena. Tal vez te suene especulativo. Pero entre los científicos de alto nivel en los laboratorios de IA, muchos ven una rápida explosión de inteligencia como sorprendentemente plausible. Pueden *verlo*. La superinteligencia es posible.

III. Los desafíos

IIIa. Corriendo hacia el grupo del billón de dólares

Se ha puesto en marcha la aceleración más extraordinaria del tecnocapital. A medida que los ingresos de la IA crezcan rápidamente, se destinarán muchos billones de dólares a la creación de GPU, centros de datos y energía antes de que finalice la década. La movilización industrial, que incluye el aumento de la producción de electricidad de Estados Unidos en un 10 por ciento, será intensa.

Verás, te dije que no se podía hacer sin convertir todo el país en una fábrica. Eso es precisamente lo que has hecho.

Niels BOHR
(a Edward Teller, al enterarse de la magnitud de la Proyecto Manhattan en 1944)

La CARRERA HACIA la AGI no solo se desarrollará en el código y detrás de las computadoras portátiles, sino que será una carrera para movilizar el poderío industrial de Estados Unidos. A diferencia de cualquier otra cosa que hayamos visto recientemente salir de Silicon Valley, la IA es un proceso industrial masivo: cada nuevo modelo requiere un nuevo clúster gigante, pronto nuevas plantas de energía gigantes y, finalmente, nuevas fábricas de chips gigantes. Las inversiones involucradas son asombrosas. Pero tras bambalinas, ya están en movimiento.

En este capítulo, te guiaré a través de los números para darte una idea de lo que esto significará:

- A medida que los ingresos de los productos de IA crecen rápidamente, alcanzando plausiblemente una tasa de ejecución anual de \$ 100 mil millones para empresas como Google o Microsoft para ~ 2026, con sistemas potentes pero anteriores a AGI,



Figura 27: El grupo de un billón de dólares. Crédito: DALLE.

eso motivará una movilización de capital cada vez mayor, y la inversión total en IA podría estar por encima de 1 billón de dólares *anuales* para 2027.

- Estamos en el camino hacia el cálculo de costos de los grupos de capacitación individuales
Miles de millones de dólares para 2028, grupos que requieren energía equivalente a la de un estado pequeño/mediano de EE.UU. y más caros que la Estación Espacial Internacional.
- A finales de la década, nos dirigimos a grupos de capacitación individuales de \$ 1T+, que requieren energía equivalente al >20% de la producción de electricidad de EE. UU. Billones de dólares de gastos de capital producirán cientos de millones de GPU por año en general.

Nvidia sorprendió al mundo cuando las ventas de sus centros de datos se dispararon de alrededor de \$ 14 mil millones *anualizados* a alrededor de \$ 90 mil millones *anualizados* en el último año. Pero eso es solo el comienzo.

Proceso de entrenamiento

Anteriormente, encontramos un crecimiento tendencial de aproximadamente ~0.5 OOMs/año de OOM = o-

Cómputo de entrenamiento de IA. Si esta tendencia continuara durante el resto de la década, ¿qué significaría eso para los grupos de capacitación más grandes?

⁴⁸ Como se mencionó anteriormente, der de magnitud, 10x = 1 orden de magnitud

Año	OOM	H100s-equivalente	Costar	Poder	Clase de referencia de potencia
2022	~GPT-4	~10k	~500 millones de dólares	~10 MW	~10,000 hogares promedio
~2024	+1 OOM	~100k	\$billions	~100MW	~100.000 viviendas
~2026	+2 OOM	~1M	\$10 millones de dólares	~1 GW	La presa Hoover, o un gran reactor nuclear
~2028	+3 OOM	~10M	\$100 de miles de millones	~10 GW	Un estado pequeño/mediano de EE.UU.
~2030	+4 OOM	~100M	\$1T+	~100GW	>20% de la producción de electricidad de EE.UU.

Esto puede parecer difícil de creer, pero parece estar sucediendo. Zuckcompró 350k H100. Amazoncompró un campus de centro de datos de 1GW junto a una planta de energía nuclear. Los rumoressugieren que se está construyendo un clúster equivalente a H100 de 1 GW y 1,4 millones (~2026 clústeres) en Kuwait. Los medios informan que se rumorea que Microsoft y OpenAI están trabajando en un clúster de \$ 100 mil millones, programado para 2028 (¡un costo comparable al de la Estación Espacial Internacional!). Y a medida que cada generación de modelos conmueve al mundo, es posible que aún se produzca una mayor aceleración.

Quizás la parte más descabellada es que la disposición a gastar ni siquiera parece ser la restricción vinculante en este momento, al menos para los clústeres de entrenamiento. Es encontrar la infraestructura en sí misma:

"¿Dónde encuentro 10GW?" (potencia para el grupo de \$ 100B +, tendencia 2028) es un tema de conversación favorito en SF. En lo que cualquier informático está pensando es en asegurar la

Tabla 4: Escalado de los grupos de entrenamiento más grandes, cálculos aproximados. Para más detalles sobre los cálculos, véase el [Apéndice](#).

energía, la tierra, los permisos, y construcción de centros de datos.⁴⁹ Si bien puede llevarle un año de esperar para obtener las GPU, los plazos de entrega para estos son mucho más largos todavía.

⁴⁹ Una incertidumbre clave es cómo. Si en lugar de necesitar esa cantidad de energía en un solo lugar, pudiéramos distribuirla entre 100 ubicaciones, sería mucho más fácil.

El clúster de un billón de dólares (+4 OOM del clúster GPT-4, el clúster de entrenamiento ~2030 en la tendencia actual) será un esfuerzo verdaderamente extraordinario. Los 100 GW de energía que requerirá equivalen al >20% de la producción de electricidad de Estados Unidos; Imagine no solo un simple almacén con GPU, sino cientos de plantas de energía. Tal vez se necesite un consorcio nacional.

(Tenga en cuenta que creo que es bastante probable que solo necesitemos un clúster de ~ \$ 100 mil millones, o menos, para AGI. El clúster de 1 billón de dólares podría ser en el que entrenaremos y ejecutaremos la superinteligencia, o lo que usaremos para AGI si AGI es más difícil de lo esperado. En cualquier caso, en un mundo post-AGI, tener la mayor cantidad de cómputo probablemente seguirá siendo realmente importante).

Proceso general

Los anteriores son solo números aproximados para los estudiantes de entrenamiento más grandes. Es probable que la inversión global sea aún mayor:

Es probable que una gran parte de las GPU se utilicen para la inferencia⁵⁰ Véase, por ejemplo, Zuckaquí; solamente

(GPUs para ejecutar los sistemas de IA de los productos), y podría haber varios jugadores con clústeres gigantes en la carrera.

Mi estimación aproximada es que 2024 ya contará con \$ 100 mil millones-

200.000 millones de dólares de inversión en IA:

- Los ingresos de los centros de datos de Nvidia pronto alcanzarán una tasa de ejecución de ~ \$ 25 mil millones / trimestre, es decir, ~ \$ 100 mil millones de gastos de capital que fluyen solo a través de Nvidia. Pero, por supuesto, Nvidia no es el único jugador (¡las TPU de Google también son geniales!), y cerca de la mitad del gasto de capital de los centros de datos está en otras cosas que los chips (sitio, edificio, refrigeración, energía, etc.).⁵¹
- Las grandes tecnológicas han aumentado drásticamente sus números de gasto de capital: Microsoft y Google probablemente ganarán \$ 50 mil millones + 52, AWS y Meta \$ 40 mil millones +, en gastos de capital este año. No todo esto es IA, pero combinados, su gasto de capital habrá crecido entre 50.000 y 100.000 millones de dólares año tras año debido al

~45k de sus H100 están en sus grupos de entrenamiento más grandes, la gran mayoría de sus 350k H100 para inferencia. Es probable que Meta tenga necesidades de inferencia más pesadas que otros jugadores, que atienden a menos clientes hasta ahora, pero a medida que los productos de IA de todos los demás escalen, espero que la inferencia se convierta en una gran mayoría de las GPU.

⁵¹ Por ejemplo, este costo total de El análisis de la propiedad estima que Alrededor del 40% del costo de un clúster grande es el

auge de la IA, e incluso entonces siguen siendo recortados. Volver a invertir en otros gastos de capital para trasladar aún más el gasto a

la IA. Además, otros proveedores de la nube, empresas (por ejemplo, Tesla está gastando 10.000 millones de dólares en IA este año) y estados-nación también están invirtiendo en IA.

GPU H100 en sí, y otro 13% va a Nvidia para las redes de Infiniband. Dicho esto, excluir el costo de capital en ese cálculo significaría que las GPU son aproximadamente el 50% del costo, y con las redes, Nvidia obtiene un poco más del 60% del costo del clúster.

⁵² Y aparentemente, a pesar de que Microsoft aumentó el gasto de capital en un 79% en comparación con hace un año en un trimestre reciente, ¡su demanda de IA en la nube aún supera la oferta!

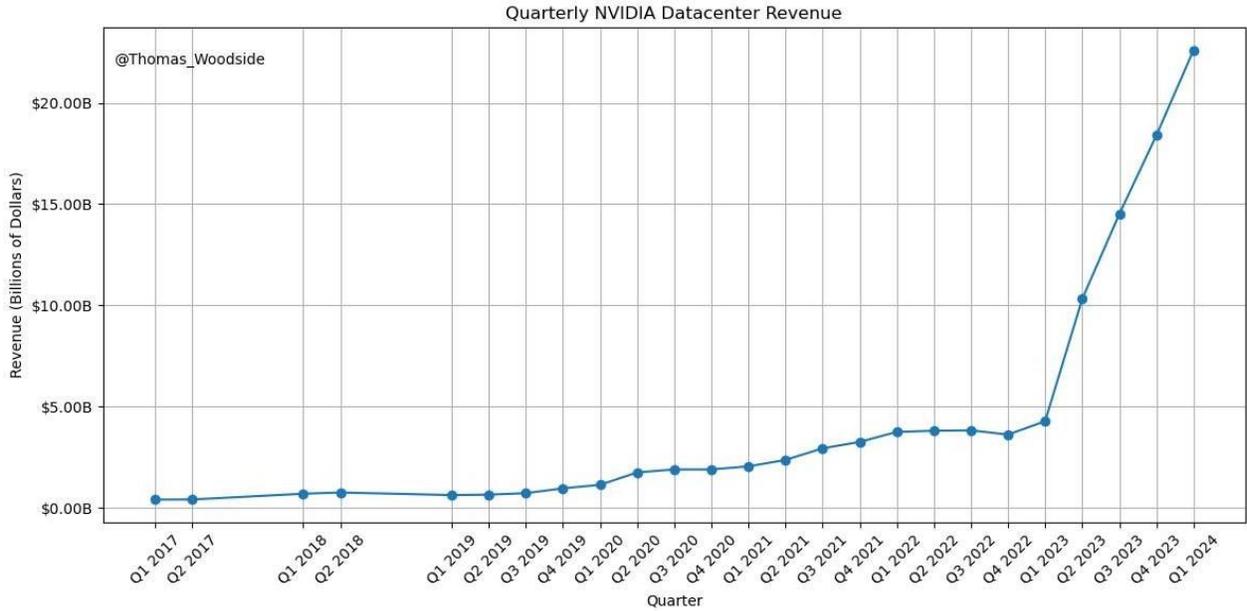


Figura 28: Ingresos trimestrales del centro de datos de Nvidia. Argumento de Thomas Woodside

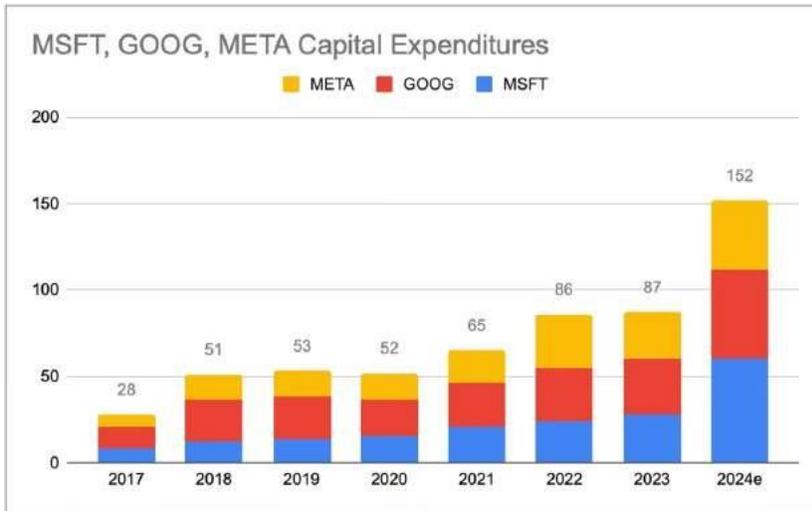


Figura 29: El gasto de capital de las grandes tecnológicas está creciendo extremadamente rápido desde que ChatGPT desató el auge de la IA. Source.

Juguemos esto hacia adelante. Mi mejor conjetura es que las inversiones informáticas generales crecerán más lentamente que las más grandes de 3 veces al año clústeres de entrenamiento, digamos 2 veces al año.⁵³ global

Y estos no son solo mis números idiosincrásicos. AMD predijoun mercado de aceleradores de IA de \$ 400 mil millones para 2027, lo que implica \$ 700 mil millones + de gasto total en IA, bastante cerca de mis números (y seguramente están mucho menos "atrapados en AGI" que yo). Sam Altman es re-

⁵³ Una fracción más grande de la GPU

Es probable que la producción vaya al mayor clúster de formación del mundo futuro que en la actualidad, por ejemplo, debido a una consolidación en unos pocos laboratorios líderes, en lugar de muchas empresas que tienen clústeres a escala de modelo de frontera. En cuanto a la potencia: Por supuesto, no todo esto será en EEUU, pero para dar una clase de referencia.

Año	Inversión anual	Envíos de aceleradores de IA (en el equivalente a H100)	Potencia como porcentaje de la producción de electricidad de EE.UU.	Chips como % de la producción actual de obleas de vanguardia de TSMC
2024	~\$150 mil millones	~5-10M	1-2%	5-10%
~2026	~\$500 mil millones	~10s de millones	5%	~25%
~2028	~\$2T	~100M	20%	~100%
~2030	~\$8T	~ 100 millones	100%	4 veces la capacidad de corriente

Tabla 5: Tendencias de juego hacia adelante en el total modelo de próxima generación. Una estimación sitúa a Microsoft en ~\$ 5 mil millones de ingresos por IA ya incrementados.

Se ha tratado de estar en conversaciones para recaudar fondos para un proyecto de "hasta 7 billones de dólares" en gastos de capital para desarrollar la capacidad de cómputo de la IA (el número fue ampliamente burlado, pero parece menos loco si se ejecutan los números aquí). De una forma u otra, esta ampliación masiva está ocurriendo.

¿Se hará? ¿Se puede hacer?

La escala DE LA inversión postulada aquí puede parecer fantástica. Pero tanto el lado de la demanda como el de la oferta parecen poder respaldar la trayectoria anterior. Los reveses económicos justifican la inversión, la escala de los gastos no tiene precedentes para una nueva tecnología de propósito general, y la movilización industrial para la energía y los chips es factible.

Ingresos de IA

Las empresas harán grandes inversiones en IA si esperan que los rendimientos económicos lo justifiquen.

Los informes sugieren que OpenAI tenía una tasa de ejecución de ingresos de \$ 1 mil millones en agosto de 2023 y una tasa de ejecución de ingresos de \$ 2 mil millones en febrero de 2024. Eso es aproximadamente una duplicación cada 6 meses. Si esa tendencia se mantiene, deberíamos ver una tasa de ejecución anual de ~ \$ 10 mil millones para fines de 2024 / principios de 2025, incluso sin descontar un aumento masivo de cualquier

inversión mundial en IA. Cálculo aproximado en el reverso del sobre. Para más detalles sobre los cálculos, véase el [Apéndice](#).

Hasta ahora, cada aumento de 10 veces en la inversión en IA parece producir los rendimientos necesarios. GPT-3.5 desató la manía de ChatGPT.

El coste estimado de 500 millones de dólares para el clúster GPT-4 se habría amortizado con los miles de millones de ingresos anuales reportados para Microsoft y OpenAI (véanse los cálculos anteriores), y un clúster de entrenamiento de "clase 2024" de miles de millones se amortizará fácilmente si los ingresos de Microsoft/OpenAI AI continúan en camino a una tasa de ejecución de 10 mil millones de dólares + de revisión. El auge está impulsado por la inversión: se necesita tiempo de un gran pedido de GPU para construir los clústeres, construir los modelos y desplegarlos, y los clústeres que se están planificando hoy en día están a muchos años vista. Pero si los rendimientos del último pedido de GPU siguen aumentando, la inversión seguirá disparándose (y superando los ingresos), invirtiendo aún más capital en una apuesta a que las próximas 10 veces seguirán dando sus frutos.

Un hito clave para los ingresos de la IA en el que me gusta pensar es: ¿cuándo alcanzará una gran empresa tecnológica (Google, Microsoft, Meta, etc.) una tasa de ejecución de ingresos de 100.000 millones de dólares de la IA (productos y API)?

Estas empresas tienen entre 100.000 y 300.000 millones de dólares de ingresos en la actualidad; Por lo tanto, 100.000 millones de dólares comenzarían a representar una fracción muy sustancial de su negocio. Extrapolar muy ingenuamente la duplicación cada 6 meses, suponiendo que alcancemos una tasa de ejecución de ingresos de \$ 10 mil millones a principios de 2025, sugiere que esto sucedería a mediados de 2026.

Eso puede parecer una exageración, pero me parece que requiere sorprendentemente poca imaginación para alcanzar ese hito. Por ejemplo, hay alrededor de **350 millones de suscriptores de pago a Microsoft Office**, ¿podría conseguir que un tercio de ellos esté dispuesto a pagar 100 dólares al mes por un complemento de IA? Para un trabajador promedio, eso es solo unas pocas horas al mes de productividad ganada; Modelos lo suficientemente poderosos como para hacer que eso justifique parezca muy factible en los próximos dos años.⁵⁴

Es difícil subestimar las reverberaciones resultantes. Esto convertiría a los productos de IA en el mayor impulsor de ingresos para las corporaciones más grandes de Estados Unidos y, con *mucho*, en su mayor área de crecimiento.

Las previsiones de crecimiento general de los ingresos de estas

empresas se dispararían. Los mercados bursátiles seguirían; Podríamos ver nuestro primer \$10 billones poco después. En este punto, las grandes tecnológicas estarían dispuestas a hacer todo lo posible, invirtiendo cada una muchos cientos de miles de millones (al menos) en una mayor ampliación de la IA. Probablemente veamos nuestro primer

⁵⁴Una gran incertidumbre para mí es cuáles son los retrasos para que la tecnología se difunda y se adopte. Creo que es plausible que los ingresos se ralenticen porque los modelos intermedios, anteriores a la AGI, requieren mucho "schlep" para integrarse adecuadamente en los flujos de trabajo de la empresa; Históricamente, se ha tardado un tiempo en cosechar plenamente las ganancias de productividad de las nuevas tecnologías de propósito. Aquí es donde entra en juego la discusión sobre el "boom sónico": a medida que "descojamos" los modelos y comienzan a parecerse más a los agentes.

en trabajadores remotos" implementarlos se vuelve mucho más fácil. En lugar de tener que rehacer por completo algún flujo de trabajo para obtener un aumento de productividad del 25% de un chatbot GPT, en su lugar

Obtendrá modelos que puede incorporar y trabajar como lo haría con un nuevo compañero de trabajo (por ejemplo, simplemente sustituir directamente a un ingeniero, en lugar de necesitar

para capacitar a los ingenieros en el uso de alguna herramienta nueva). O, en el extremo, y más adelante: no necesitarás rediseñar completamente una fábrica para trabajar con alguna herramienta nueva, solo traerás los robots humanoides.

Dicho esto, esto puede conducir a cierta discontinuidad en el valor económico y los ingresos generados, dependiendo de la rapidez con la que podamos "destrabar" los modelos.

venta de bonos corporativos por cientos de miles de millones de dólares.⁵⁵ ⁵⁵ ¿Qué pasará con las tasas de interés?

será interesante... véase Tyler Cowen aquí; Chow, Mazlish y Halperin aquí.

Más allá de los 100.000 millones de dólares, cada vez es más difícil ver los contornos. Pero si realmente estamos en el camino hacia el AGI, los rendimientos estarán ahí. A los trabajadores de cuello blanco se les pagan decenas de billones de dólares en salarios anuales en todo el mundo; un trabajador remoto que automatiza incluso una fracción de los trabajos cognitivos/de cuello blanco (imagínese, por ejemplo, un codificador de IA verdaderamente automatizado) pagaría por el grupo de un billón de dólares. Al menos, la importancia de la seguridad nacional bien podría motivar un proyecto de gobierno, agrupando los recursos de la nación en la carrera hacia la AGI (más adelante).

Precedentes históricos

1 billón de dólares al año de inversión anual total en IA para 2027 parece exagerado. Pero vale la pena echar un vistazo a otras clases de referencia histórica:

- En sus años de mayor financiación, los programas Manhattan y Apolo representaron cada uno el 0,4% del PIB, o ~ \$ 100 mil millones anuales en la actualidad (¡sorprendentemente pequeño!). A 1 billón de dólares al año, la inversión en IA supondría alrededor del 3% del PIB.
- Entre 1996 y 2001, las empresas de telecomunicaciones invirtieron casi 1 billón de dólares de hoy en la construcción de infraestructura de Internet.
- De 1841 a 1850, las inversiones ferroviarias privadas británicas sumaron un acumulado de ~ 40% del PIB británico en ese momento. Una fracción similar del PIB de EE.UU. equivaldría a ~\$ 11T en una década.
- Se están gastando muchos billones en la transición ecológica.
- Las economías de rápido crecimiento a menudo gastan una fracción alta por ejemplo, China ha gastado más del 40% de su PIB en inversión durante dos décadas (equivalente a 11 billones de dólares anuales dado el PIB de Estados Unidos).
- En las circunstancias históricamente más exigentes de seguridad nacional, en tiempos de guerra, el endeudamiento para financiar el esfuerzo nacional a menudo ha representado enormes

fracciones del PIB. Durante la Primera Guerra Mundial, el

Reino Unido y Francia, and Germany prestados El 100% de sus PIB, mientras que el NOS se endeudó más del 20%; durante la Segunda Guerra Mundial, el Reino Unido y Japón tomó prestado más del 100% de su PIB, mientras que el NOS prestó más del 60% del PIB (equivalente a más de 17 billones de dólares en la actualidad).

1 billón de dólares al año de inversión total en IA para 2027 sería espectacular, una de las mayores acumulaciones de capital de la historia, pero no carecería de precedentes. Y un entrenamiento individual de un billón de dólares para finales de la década parece estar sobre la mesa.⁵⁶ liderara

Poder

Probablemente, la mayor limitación en el lado de la oferta será la energía. A escalas a corto plazo (1 GW/2026 y especialmente 10 GW/2028), la energía se ha convertido en una limitación vinculante: simplemente no hay mucha capacidad ociosa y los contratos de energía suelen estar bloqueados a largo plazo. Y construir, por ejemplo, una nueva planta de energía nuclear de gigavatios lleva una década. (No sabré cuándo empezaremos a ver cosas como que las empresas tecnológicas compren empresas de fundición de aluminio por su potencia de gigavatios contratos.⁵⁷)

⁵⁶ Y, más lejos, pero si AGI realmente

a aumentos sustanciales en el crecimiento económico, \$10T+ anuales comenzarían siendo plausible, ya que la clase de referencia son las tasas de inversión de los países durante los períodos de alto crecimiento.

⁵⁷ "Desde 2011, la fundición Alouette utiliza 930 MW de electricidad a su máxima capacidad de producción".

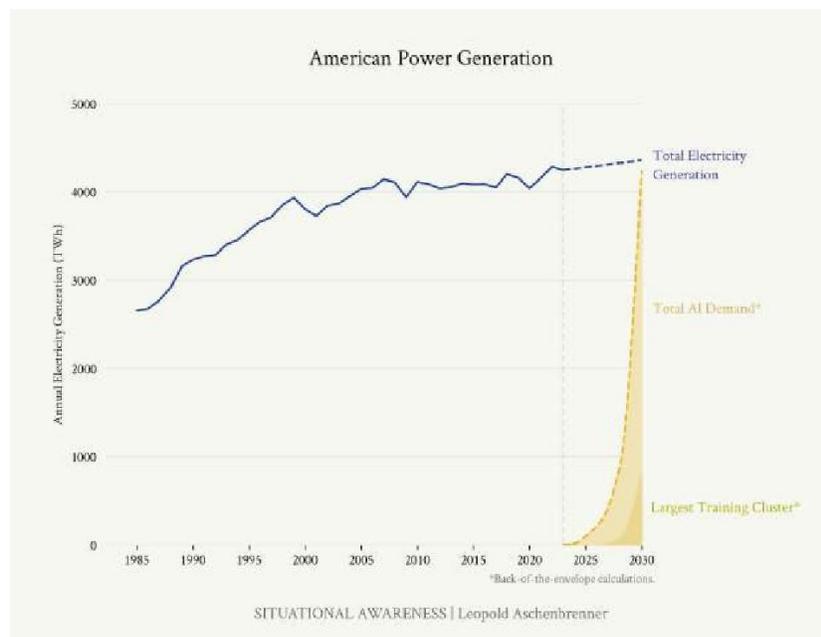


Figura 30: Comparación de las tendencias de la producción total de electricidad de EE. UU. con nuestras estimaciones preliminares sobre la demanda de electricidad de IA.

La generación total de electricidad en EE.UU. apenas ha crecido un 5% en los últimos años

década. Las empresas de servicios públicos están empezando a entusiasmarse con la IA (en lugar de un crecimiento del 2,6% en los próximos 5 años, ¡ahora estiman un 4,7%!). Pero apenas están valorando lo que viene. El clúster de 100 GW de un billón de dólares por sí solo requeriría ~ 20% de la generación eléctrica actual de EE. UU. en 6 años; Junto con una gran capacidad de inferencia, la demanda será múltiplo mayor.

Para la mayoría, esto parece completamente fuera de discusión. Algunos están apostando por las autocracias de Oriente Medio, que han estado ofreciendo un poder ilimitado y grupos gigantes para conseguir que sus gobernantes tengan un asiento en la mesa de la AGI.

Pero es totalmente posible hacer esto en los Estados Unidos: tenemos abundante gas natural.⁵⁸

⁵⁸ Gracias a Austin Vernon (correspondencia privada) para ayudar con estas Estimaciones.

- Alimentar un clúster de 10 GW tomaría solo un pequeño porcentaje de la producción de gas natural de EE. UU. y podría hacerse rápidamente.
- Incluso el clúster de 100 GW es sorprendentemente factible.
 - En este momento, el esquisto Marcellus/Utica (alrededor de Pensilvania) por sí solo está produciendo alrededor de 36 mil millones de pies cúbicos por día de gas; eso sería suficiente para generar poco menos de 150 GW continuamente con generadores (y las centrales eléctricas de ciclo combinado podrían producir 250 GW debido a su mayor eficiencia).
 - Se necesitarían alrededor de ~ 1200 nuevos pozos para los 100 GW de cluster.⁵⁹ Cada equipo puede perforar aproximadamente 3 pozos por mes, es decir, 40 nuevos pozos producen alrededor de 0,01 BCF al día. ⁵⁹ Los (el número actual de plataformas en el Marcellus) podría construir aumentar la base de producción para 100 GW en menos de un año.⁶⁰ ⁶⁰ Cada pozo produce ~20 BCF más Su vida útil, lo que significa que dos nuevos pozos al mes reemplazarían las reservas agotadas. Es decir, solo necesitaría un equipo para mantener la producción.
 - En términos más generales, la producción de gas natural de EE.UU. se ha más que duplicado en una década; el simple hecho de continuar esa tendencia podría Impulse varios centros de datos de billones de dólares.⁶² ⁶¹ Aunque sería más eficiente agregar menos plataformas y construir en un período de tiempo más largo que 10 meses. ⁶² Un pie cúbico de gas natural Genera alrededor de 0,13 kWh. Producción de

- La parte más difícil sería construir suficientes generadores/turbinas; Esto no sería trivial, pero parece factible con

gas de esquisto
fue de aproximadamente ~ 70 mil millones de pies cúbicos por

día en los EE. UU. en 2020. Supongamos que volvemos a duplicar la producción y que toda la capacidad adicional se destina a clústeres informáticos. Eso es 3322 TWh/año de electricidad, o suficiente para casi 4 clústeres de 100 GW.

alrededor de \$ 100 mil millones de capex⁶³ para 100 GW de energía de gas natural ⁶³Los costos de inversión de capital para la energía de gas natural plantas. Las centrales de ciclo combinado pueden construirse en unos dos años; El cronograma para los generadores sería aún más corto.⁶⁴ chips de IA (incluida la

Las barreras para incluso billones de dólares de construcción de centros de datos en los EE. UU. son totalmente autosuficientes. Los compromisos climáticos bien intencionados pero rígidos (no solo por parte del gobierno, sino también por parte de Microsoft, Google, Amazon, etc.) se interponen en el camino de la solución obvia y rápida. Por lo menos, incluso si no hacemos gas natural, una amplia agenda de desregulación desbloquearía los megaproyectos solares, baterías, SMR y geotérmicos. Los permisos, la regulación de los servicios públicos, la regulación de las líneas de transmisión por parte de la FERC y la revisión ambiental de la NEPA hacen que las cosas que deberían llevar unos años tarden una década o más. No tenemos ese tipo de tiempo.

Vamos a llevar los centros de datos de AGI a Oriente Medio, bajo el control de autócratas brutales y caprichosos. Yo también preferiría la energía limpia, pero esto es simplemente demasiado importante para la seguridad nacional de Estados Unidos. Necesitaremos un nuevo nivel de determinación para que esto suceda. La restricción de poder puede, debe y será resuelta.

Patatas fritas

Si bien los chips suelen ser lo que viene a la mente cuando la gente piensa en las limitaciones de suministro de IA, es probable que sean una restricción menor que la energía. La producción mundial de chips de IA sigue siendo un porcentaje bastante pequeño de la producción de vanguardia de TSMC, probablemente menos del 10%. Hay mucho espacio para crecer a través de que la IA se convierta en una mayor parte de la producción de TSMC.

De hecho, la producción de chips de IA en 2024 (~5-10 millones equivalentes a H100) ya sería casi suficiente para el grupo de 100 mil millones de dólares (si todos se desviarán a un clúster). Desde un punto de vista de fabricación puramente lógica, ~ el 100% de la producción de TSMC durante un año ya podría soportar el clúster de un billón de dólares (nuevamente si todos los chips fueran a un centro de datos).⁶⁵ Por supuesto, no toda TSMC podrá ser desviada a la IA, y no toda la producción de chips de IA durante un año será para un clúster de entrenamiento. Demanda total de

Las plantas parecen estar por debajo de los \$ 1000 por kW, lo que significa que el gasto de capital es de 100 GW de las centrales eléctricas de gas natural serían de unos 100.000 millones de dólares.

⁶⁴ La energía solar y las baterías no son una alternativa totalmente descabellada, pero parece más dura que el gas natural. Aprecié el cálculo de Casey Handmer de embaldosar la Tierra en paneles solares:

"Con las GPU actuales, la computación del centro de datos solar global es equivalente a ~ 150 mil millones de humanos, aunque si nuestras computadoras eventualmente pueden igualar la eficiencia [del cerebro humano], podríamos admitir más de 5 cuatrillones de almas de IA".

⁶⁵ Esto plantea la interesante pregunta de por qué las necesidades de energía están aumentando tanto antes de que la producción de fábricas de chips comience a verse realmente limitada.

Una respuesta simple es que, si bien los centros de datos funcionan continuamente a una potencia cercana a la máxima, la mayoría de los chips que se producen actualmente están inactivos la mayor parte del tiempo. En la actualidad, los teléfonos inteligentes representan cerca de la mitad de la demanda de chips líderes, pero utilizan mucho menos energía por área de oblea (transistores comerciales para operaciones en serie y eficiencia energética) y tienen una baja utilización ya que los teléfonos inteligentes están en su mayoría inactivos. La revolución de la IA significa trabajar mucho más duro con nuestros transistores, dedicándolos todos a centros de datos de IA de alto rendimiento y en constante funcionamiento en lugar de dispositivos inactivos, alimentados por baterías y que ahorran energía. HT Carl Shulman para este punto.

inferencia y múltiples jugadores) para 2030 será un múltiplo de la capacidad total actual de chips lógicos de vanguardia de TSMC, solo para

IA. TSMC ~ **Doblado**⁶⁶ en los últimos 5 años; Es probable que tengan que hacerlo ⁶⁶ (Uso de los ingresos como proxy).

ir ~ al menos dos veces más rápido en su ritmo de expansión para satisfacer la demanda de chips de IA. Se necesitarían nuevas inversiones masivas en fábricas.

Incluso si las fábricas de lógica en bruto no serán la restricción, el empaquetado avanzado de chip en oblea en sustrato (CoWoS) (que conecta los chips a la memoria, también fabricados por TSMC, Intel y otros) y la memoria HBM (cuya demanda **es enorme**) ya son cuellos de botella clave para el actual escalado de GPU de IA; estos están más especializados para la IA, a diferencia de los chips lógicos puros, por lo que hay menos capacidad preexistente. A corto plazo, estas serán las principales limitaciones para producir más GPU, y estas serán las enormes limitaciones a medida que la IA se amplíe. Aun así, estos son comparativamente "fáciles" de escalar; ha sido increíble ver a TSMC construir literalmente fábricas "greenfield" (es decir, instalaciones completamente nuevas desde cero) para aumentar masivamente la producción de CoWoS este año (y Nvidia incluso está empezando a encontrar alternativas de **CoWoS** para evitar la escasez).

Un nuevo TSMC **Gigafab** (una **maravilla tecnológica**) cuesta alrededor de \$ 20 mil millones en gastos de capital y produce 100 mil inicios de obleas al mes. Para cientos de millones de GPU de IA al año para finales de la década, TSMC necesitaría construir docenas de ellas, así como una enorme acumulación de memoria, empaquetado avanzado, redes, etc., lo que supondrá una fracción importante del gasto de capital. Podría añadir hasta más de 1 billón de dólares de gasto de capital. Será intenso, pero factible. (Tal vez el mayor obstáculo no sea la viabilidad, pero TSMC ni siquiera lo intenta: ¡TSMC aún no parece tener una píldora de escalado de IA! Creen que la IA "sólo" crecerá a una **tasas compuesta anual glacial del 50%**).

Los esfuerzos recientes del Gobierno de los Estados Unidos, como la Ley CHIPS, han estado tratando de llevar a cabo una mayor producción de chips de IA en los EE. UU. (como seguro en caso de la contingencia de Taiwán). Si bien sería bueno deslocalizar una mayor producción de chips de IA a los EE. UU., Es menos crítico que tener el centro de datos real (en el que vive el AGI) en los EE. UU. Si tener producción de chips en el extranjero es como tener depósitos de uranio en el extranjero, tener el centro de datos AGI

en el extranjero es como tener las armas nucleares literalmente
construidas y almacenadas en el extranjero. Dada la disfunción

y el costo que hemos visto de la construcción de fábricas en los EE. UU. en la práctica, supongo que deberíamos priorizar los centros de datos en los EE. UU. mientras apostamos más fuertemente por aliados democráticos como Japón y Corea del Sur para proyectos fabulosos: las construcciones de fábricas allí **parecen mucho más** funcionales.

Los cúmulos de la democracia

Antes DE QUE TERMINE EL EVENTO , se habrán construido muchos billones de dólares en clústeres de cómputo. La única pregunta es si se construirán en Estados Unidos. Se rumorea que algunos están apostando por construirlos en otros lugares, especialmente en Oriente Medio. ¿Realmente queremos que la infraestructura del Proyecto Manhattan sea controlada por alguna caprichosa dictadura de Oriente Medio?

Los grupos que se están planificando hoy en día bien pueden ser los centros en los que se entrenan y ejecutan la AGI y la superinteligencia, no solo los "clústeres de productos de gran tecnología". El interés nacional exige que estos se construyan en Estados Unidos (o en aliados democráticos cercanos). Cualquier otra cosa crea un riesgo de seguridad irreversible:

corre el riesgo de que los pesos AGI sean robados⁶⁷ (y tal vez enviado

en- **a China**); se corre el riesgo de que estas dictaduras se apoderen físicamente de los centros de datos (para construir y dirigir AGI ellos mismos) cuando la carrera de AGI se calienta; o incluso si estas amenazas solo se ejercen implícitamente, pone a AGI y superinteligencia a los caprichos desagradables del dictador. Estados Unidos lamentó profundamente su dependencia energética de Oriente Medio en los años 70, y trabajamos muy duro para librarnos de sus manos. No podemos volver a cometer el mismo error.

Los clústeres se pueden construir en los EE.UU., y tenemos que ponernos las pilas para asegurarnos de que esto suceda en los EE.UU. La seguridad nacional de Estados Unidos debe ser lo primero, antes que el atractivo del dinero que fluye libremente en el Medio Oriente, la regulación arcana o incluso, sí, los admirables compromisos climáticos. Nos enfrentamos a una verdadera competencia sistémica: ¿puede hacerse la movilización industrial necesaria sólo en la "parte superior" de la

⁶⁷ Es mucho más fácil hacer un canal lateral

¿Autocracias "abajo"? Si las empresas estadounidenses se liberan, América puede construir como ninguna otra (al menos en los estados rojos). Estar dispuesto a usar gas natural, o al menos una agenda desreguladora de base amplia: exenciones de la NEPA, fijación de la FERC y

¡Tachuelas para exfiltrar pesas con acceso físico!

Los permisos de transmisión a nivel federal, la regulación de los servicios públicos, el uso de las autoridades federales para desbloquear tierras y derechos de paso, es una prioridad de seguridad nacional.

En CUALQUIER caso: lo exponencial está en pleno apogeo ahora.

En los "viejos tiempos", cuando la AGI todavía era una mala palabra, algunos colegas y yo solíamos hacer modelos económicos teóricos de cómo podría ser el camino hacia la AGI. Una característica de estos modelos solían ser un hipotético momento de "despertar de la IA", cuando el mundo comenzó a darse cuenta de lo poderosos que podían ser estos modelos y comenzó a aumentar rápidamente sus inversiones, lo que culminó en múltiples porcentajes del PIB hacia las mayores carreras de entrenamiento.

Parecía lejano entonces, pero ese momento ha llegado. 2023 fue "IA Despierta".⁶⁸ Detrás de las escenas, la tecnología más asombrosa se ha puesto en marcha la aceleración del

⁶⁸ Recuerdo claramente haber escrito "EL DESPEGUE HA COMENZADO" en mi pizarra en marzo de 2023.

capital. Prepárate para las fuerzas G.

(Lo que todo esto significa para NVDA/TSM/etc lo dejo como un ejercicio para el lector. Pista: Aquellos con conciencia situacional compraron mucho más bajo que usted, pero todavía no está ni cerca de su precio total).⁶⁹
venta parecen

⁶⁹ Los principales analistas del lado de la

para asumir solo un crecimiento interanual del 10-20% en los ingresos de Nvidia desde el año fiscal 24 hasta el año fiscal 25, tal vez entre 120 mil millones y 130 mil millones de dólares en el año fiscal 25 (o al menos lo hizo hasta hace muy poco). ¡Loco! Ha sido bastante obvio durante un tiempo que Nvidia va a obtener más de \$ 200 mil millones de ingresos en CY25.

IIIb. Bloqueo de los laboratorios: seguridad para AGI

Los principales laboratorios de IA del país tratan la seguridad como una ocurrencia tardía. Actualmente, básicamente están entregando los secretos clave de AGI al PCCh en bandeja de plata. Asegurar los secretos y pesos de AGI contra la amenaza de los actores estatales será un esfuerzo inmenso, y no vamos por buen camino.

Se reunieron por la noche en la oficina de Wigner. "Szilard esbozó los datos de Columbia", informa Wheeler, "y las indicaciones preliminares de que al menos dos neutrones secundarios emergen de cada fisión inducida por neutrones. ¿No significaba esto que un explosivo nuclear era ciertamente posible?" No necesariamente, replicó Bohr.

"Intentamos convencerlo", escribe Teller, "de que debíamos seguir adelante con la investigación de la fisión, pero que no debíamos publicar los resultados. Deberíamos mantener los resultados en secreto, no sea que los nazis se enteren de ellos y produzcan explosiones nucleares primero".

"Bohr insistió en que nunca tendríamos éxito en la producción de energía nuclear y también insistió en que el secreto nunca debe introducirse en la física".

*La fabricación de la bomba atómica, p.
430*

En EL curso actual, los principales laboratorios chinos de AGI no estarán en Pekín o Shanghái, sino en San Francisco y Londres. Dentro de unos años, quedará claro que los secretos de la AGI son los secretos de defensa nacional más importantes de los Estados Unidos, y merecen un tratamiento a la par con los planos de los bombarderos B-21 o los submarinos de la clase Columbia, por no hablar de los proverbiales

"secretos nucleares", pero hoy en día, los estamos tratando de la misma manera que lo haríamos con un software SaaS aleatorio. A este ritmo, básicamente estamos entregando superinteligencia al PCCh.

Todos los billones que invertiremos, la movilización del poderío industrial estadounidense, los esfuerzos de nuestras mentes más brillantes, nada de eso importa si China u otros pueden simplemente robar los *pesos del modelo* (todo lo que es un modelo de IA terminado, todo lo que será AGI, es un gran archivo en una computadora) o *los secretos algorítmicos clave* (los avances técnicos clave necesarios para construir AGI).

Los principales laboratorios de IA de Estados Unidos se autoproclaman que están construyendo AGI: creen que la tecnología que están construyendo será, antes de que termine la década, el arma más poderosa que Estados Unidos haya construido. Pero no lo tratan como tal. Miden sus esfuerzos de seguridad en función de "nuevas empresas tecnológicas aleatorias", no de "proyectos clave de defensa nacional". A medida que la carrera de AGI se intensifique, a medida que quede claro que la superinteligencia será absolutamente decisiva en la competencia militar internacional, tendremos que enfrentarnos a toda la fuerza del espionaje extranjero. En la actualidad, los laboratorios apenas pueden defenderse de los scriptkiddies, y mucho menos tener una "seguridad a prueba de Corea del Norte", y mucho menos estar preparados para enfrentarse al Ministerio de Seguridad del Estado chino que utiliza toda su fuerza.

Y esto no solo importará en los próximos años. Claro, a quién le importa si las pesas GPT-4 son robadas: lo que realmente importa en términos de seguridad de peso es que podemos asegurar las pesas AGI en el futuro, por lo que tenemos algunos años, podría decirse. (Aunque si estamos construyendo AGI en 2027, ¡realmente tenemos que movernos!) Pero el Los laboratorios de IA están desarrollando los *secretos algorítmicos* (los avances técnicos clave, los planos, por así decirlo) para el AGI *en este momento* (en particular, el "próximo paradigma" de RL/auto-reproducción/datos sintéticos/etc. después de los LLM para superar el muro de datos). La seguridad de nivel AGI para los secretos algorítmicos es necesaria años antes de la seguridad de nivel AGI para los pesos. Estos avances algorítmicos importarán más que un clúster 10 o 100 veces más grande en unos pocos años, esto es mucho más importante que los controles de exportación en cómputo, que el Gobierno de los Estados

Unidos ha estado (¡proféticamente!) persiguiendo intensamente. En este momento, ni siquiera necesitas montar una dramática operación de espionaje para robar estos secretos: solo ve a cualquier fiesta de ciencia ficción o mira a través de las ventanas de la oficina.

Nuestro fracaso de hoy será irreversible pronto: en los próximos 12 a 24 meses, filtraremos avances clave de AGI al PCCh. Será el mayor arrepentimiento del establishment de seguridad nacional antes de que termine la década.

La **preservación del mundo libre** frente a los estados autoritarios está en juego, y una ventaja saludable será el amortiguador necesario que nos dé margen para que la seguridad de la IA también sea correcta. Estados Unidos tiene una ventaja en la carrera por el AGI. Pero renunciaremos a este liderazgo si no nos tomamos en serio la seguridad muy pronto. Ponerse manos a la obra, ahora, es tal vez incluso lo más importante que tenemos que hacer hoy para garantizar que la AGI vaya bien.

Subestimar a los actores estatales bajo su propio riesgo

Demasiadas personas inteligentes subestiman el espionaje.

Las capacidades de los Estados y sus agencias de inteligencia son extremadamente formidables. Incluso en tiempos normales, sin una carrera de AGI total (y por lo poco que sabemos públicamente), los estados-nación (o actores menos avanzados) han sido capaces de:

- **Zer-o-click** hackea cualquier iPhone y Mac deseado con solo un número de teléfono,
- Infiltrarse en un **programa de armas atómicas con espacio abierto**,
- **Modificar el código fuente de Google**,
- Encontrar **Docenas de días** cero a año en que se tomó **Promedio de 7 años** para detectar,
- **Spearfish** grandes **empresas tecnológicas**,
- **Instalar** **loggers** en los dispositivos de los empleados,
- **Insertar** **trampillas** en los esquemas de **encriptación**,
- **Robar** información a través de **emanaciones electromagnéticas o vibración**,

- Usa solo el ruido de tu computadora para determinar dónde te encuentras en el mapa de un videojuego orobar una contraseña,
 - Obtener acceso directo a sistemas sensibles como las centrales nucleares,
 - Exfiltrar 22 millones de archivos de autorización de seguridad del Gobierno de los Estados Unidos,
 - Exponga el archivo Información financiera de 110 millones de clientes por plantar vulnerabilidades en los sistemas de climatización,
 - Comprometer las cadenas de suministro de hardware informático a gran escala,
 - Introducir código malicioso en las actualizaciones de las dependencias de software utilizadas por las principales empresas tecnológicas y el Gobierno de los Estados Unidos
- emplear medios extraordinarios y pagar cualquier costo para infiltrarse en los laboratorios de IA.
- . . . por no hablar de plantar espías o seducir, engatusar o amenazar a los empleados (lo que ocurre eficazmente a gran escala, pero es menos público)
 - . . . por no hablar de las operaciones de las fuerzas especiales y similares (cuando las cosas se ponen realmente calientes).

Para una muestra más profunda de lo que estamos tratando cuando nos enfrentamos a las agencias de inteligencia, recomiendo encarecidamente *Inside the Aquarium*, un libro escrito por un miembro soviético del GRU (inteligencia militar).⁷⁰

China ya se dedica a un espionaje industrial generalizado; el director del FBI declaró que la República Popular China tiene una operación de piratería informática mayor que "todas las naciones importantes juntas". Y hace apenas un par de meses, el Fiscal General anunció la detención de un ciudadano chino que había robado un código clave de IA de Google para llevárselo a la República Popular China (en 2022/23, y probablemente solo el punta del iceberg).⁷¹

Pero eso es solo el comienzo. Debemos estar preparados para que nuestros adversarios "despierten a AGI" en los próximos años. La IA se convertirá en la prioridad #1 de todas las agencias de inteligencia del mundo. En esa situación, estarían dispuestos a

⁷⁰ Un spoiler, a modo de muestra: al graduarse de su ~academia de espionaje, antes de ser enviados al extranjero, los aspirantes a espías tenían que demostrar sus habilidades en el país: tenían que adquirir información secreta de un científico soviético. El castigo por revelar secretos de Estado, por supuesto, era la muerte. Es decir: graduarse de la ~academia de espionaje significaba elegir a un compatriota para condenarlo a muerte. HT Ilya Sutskever por la recomendación del libro.

⁷¹ Por cierto, la acusación proporciona una gran ilustración de lo fácil que es evadir la seguridad incluso en Google, que probablemente tenga la mejor seguridad de todos los laboratorios de IA (dada su capacidad para apoyarse en la inversión de décadas de Google en infraestructura de seguridad). ¡Todo lo que se necesitó para robar el código, sin ser detectado, fue pegar el código en Apple Notes y luego exportarlo a pdf!

"DING exfiltró estos archivos copiando datos de los archivos fuente de Google en la aplicación Apple Notes en su computadora portátil MacBook emitida por Google. A continuación, DING convirtió las notas de Apple en archivos PDF. y los cargó desde la red de Google en la cuenta DING 1. Este método ayudó a DING a evadir la detección inmediata". (De la [acusación](#)).

Solo lo atraparon porque hizo un montón de otras cosas estúpidas, como inmediatamente comenzaron nuevas empresas prominentes en China, lo que hizo que la gente sospechara (y luego incluso regresó a los EE. UU.).

El modelo de amenazas

Hay DOS activos CLAVE que debemos proteger: los pesos de los modelos (especialmente a medida que nos acercamos a AGI, pero que requiere años de preparación y práctica para hacerlo bien) y los secretos algorítmicos (a partir de ayer).

Pesos del modelo

Un modelo de IA no es más que un gran archivo de números en un servidor. Esto puede ser robado. Todo lo que se necesita de un adversario para igualar sus billones de dólares y sus mentes más inteligentes y sus décadas de trabajo es robar este archivo. (Imagínese si los nazis hubieran conseguido un duplicado exacto de cada bomba atómica fabricada en Los Álamos).

Si no podemos mantener seguros los pesos de los modelos, solo estamos construyendo AGI para el PCCh (y, dada la trayectoria actual de la seguridad de los laboratorios de IA, incluso para Corea del Norte).

Incluso además de la competencia nacional, asegurar los pesos de los modelos también es fundamental para prevenir catástrofes de IA. Todas nuestras medidas de protección y de protección no sirven de nada si un mal actor (por ejemplo, un terrorista o un estado canalla) puede robar el modelo y hacerlo que quiera con él, eludiendo cualquier capa de seguridad. Cualquier nueva ADM que la superinteligencia pudiera inventar proliferaría rápidamente en docenas de estados canallas. Además, la seguridad es la primera línea de defensa contra el descontrol o la desalineación

También sistemas de IA (¿qué estúpidos nos sentiríamos si no contuviéramos la superinteligencia rebelde porque no la construimos y probamos primero en un grupo aislado?).

Asegurar los pesos de los modelos no importa mucho en este momento: robar GPT-4 sin la receta subyacente no hace mucho por el PCCh. Pero *realmente* importará en unos años, una vez que tengamos AGI, sistemas que son realmente increíblemente poderosos.

Quizás el escenario que más me quita el sueño es si China u otro adversario es capaz de robar los pesos de los modelos de investigación de IA automatizada en la cúspide de una explosión de inteligencia. China podría utilizarlos inmediatamente para automatizar la investigación de la IA por sí misma

(incluso si previamente habían estado muy atrasados)— y lanzan su propia explosión de inteligencia. Eso sería todo lo que necesitan para automatizar la investigación de IA y construir superinteligencia. Cualquier pista que tuviera Estados Unidos se desvanecería.

Además, esto nos pondría inmediatamente en una carrera existencial; Cualquier margen para garantizar la seguridad de la superinteligencia desaparecería. Es muy posible que el PCCh intente atravesar una explosión de inteligencia lo más rápido posible —incluso meses de ventaja en la superinteligencia podrían significar una ventaja militar decisiva— en el proceso, saltándose todas las precauciones de seguridad que cualquier esfuerzo responsable de AGI de EE. UU. esperaría tomar. También tendríamos que correr a través de la explosión de inteligencia para evitar el dominio total del PCCh. Incluso si EE.UU. apenas logra salir adelante al final, la pérdida de margen significaría tener que correr riesgos enormes en la seguridad de la IA.

Hoy en día estamos a kilómetros de distancia para tener la seguridad suficiente para proteger las pesas. Google DeepMind (quizás el laboratorio de IA que tiene la mejor seguridad de cualquiera de ellos, dada la infraestructura de Google) lo admite, al menos directamente. Su [Marco de Seguridad Fronteriza](#) describe los niveles de seguridad 0, 1, 2, 3 y 4 (~ 1,5 es lo que necesitaría para defenderse de grupos terroristas o criminales con buenos recursos, 3 es lo que necesitaría para defenderse de las Corea del Norte del mundo, y 4 es lo que necesitaría para tener incluso una oportunidad de defenderse contra los esfuerzos prioritarios de la los actores estatales más capaces).⁷² Admiten estar en el nivel 0 corresponsales

(solo las medidas más banales y básicas). Si tuviéramos AGI y superinteligencia pronto, ¡literalmente se los entregaríamos a los grupos terroristas y a todos los dictadores locos que existen!

De manera crítica, el desarrollo de la infraestructura para la seguridad del peso probablemente lleve muchos años de plazos de entrega, si pensamos en AGI en ~ 3-4 años es una posibilidad real y necesitamos una seguridad de peso a prueba de estado entonces, necesitamos lanzar el esfuerzo de choque *ahora*.

Asegurar los pesos requerirá innovaciones en el hardware y un diseño de clúster radicalmente diferente; Y la seguridad en este nivel no se puede alcanzar de la noche a la mañana, sino que requiere ciclos de iteración.

⁷² Sobre la base de sus supuestos

Si no nos preparamos a tiempo, nuestra situación será grave. Estaremos en la cúspide de la superinteligencia, pero a años de distancia de la

De acuerdo con el informe de seguridad de peso de RAND, el informe de seguridad L1-L5

seguridad necesaria. Nuestra opción será seguir adelante, pero entregar directamente la superinteligencia al PCCh, con la carrera existencial a través de la explosión de inteligencia que eso implica, o esperar hasta que se complete el programa de choque de seguridad, arriesgándonos a perder nuestra ventaja.

Secretos algorítmicos

Si bien la gente está empezando a apreciar (aunque no necesariamente a implementar) la necesidad de la seguridad del peso, podría decirse que aún más importante en este momento, y *muy* subestimada, es la seguridad de los secretos algorítmicos.

Una forma de pensar en esto es que robar los secretos algorítmicos valdrá la pena tener un clúster 10 veces o más grande para la República Popular China:

- Como se explica en Recuento de las OOM, el progreso algorítmico es probablemente tan importante como escalar el proceso al progreso de la IA. Dada la tendencia de referencia de $\sim 0,5$ OOM de eficiencia de cómputo al año (+ ganancias algorítmicas adicionales de "descojeación" en la parte superior), deberíamos esperar múltiples OOM de secretos algorítmicos entre ahora y AGI. Por defecto, espero que los laboratorios estadounidenses estén años por delante; Si pueden defender sus secretos, esto podría valer fácilmente un cálculo de 10x-100x.
 - (Tenga en cuenta que estamos dispuestos a incurrir en costos de miles de millones de dólares para los inversores estadounidenses al exportar chips de control de Nvidia, tal vez un aumento de 3 veces en el costo de cómputo para los laboratorios chinos, ¡pero estamos filtrando secretos algorítmicos de 3 veces por todas partes!)
- Quizás lo más importante es *que podemos estar desarrollando los avances clave del paradigma para AGI en este momento*. Como ya se ha comentado anteriormente, el simple hecho de ampliar los modelos actuales se topará con un muro: el muro de los datos. Incluso con mucha más computación, no será posible hacer un modelo mejor. Los laboratorios de IA de frontera están trabajando intensamente en lo que viene a continuación, desde RL hasta datos sintéticos. Probablemente descubrirán algunas cosas locas, esencialmente, el equivalente de "autojuego de AlphaGo" para la inteligencia general. Sus invenciones serán tan importantes como lo fue la invención del paradigma LLM hace varios años

y serán la clave para construir sistemas que vayan mucho más allá del nivel humano. Todavía tenemos la oportunidad de negar a China estos avances algorítmicos clave, sin los cuales estarían atrapados en la pared de datos. Pero sin una mejor seguridad en los próximos 12 a 24 meses, es muy posible que suministremos irreversiblemente a China estos avances clave en materia de AGI.

- Es fácil subestimar lo importante que serán los secretos algorítmicos de borde, porque hasta ~hace un par de años, todo se publicaba. La idea básica estaba ahí: escalar Transformers en el texto de Internet. Había muchos detalles algorítmicos y eficiencias: leyes de escalado de Chinchilla, MoE, etc. Por lo tanto, los modelos de código abierto de hoy en día son bastante buenos, y Un puñado de empresas tienen modelos bastante buenos (en su mayoría dependiendo de la cantidad de dólares que recaudaron y del tamaño de sus clústeres). Pero es probable que esto cambie drásticamente en los próximos dos años. Básicamente, todo el progreso algorítmico de frontera ocurre en los laboratorios en estos días (la academia es sorprendentemente irrelevante), y los laboratorios líderes han dejado de publicar sus avances. Deberíamos esperar mucha más divergencia en el futuro: entre laboratorios, entre países y entre la frontera propietaria y los modelos de código abierto. Unos pocos laboratorios americanos estarán muy por delante, un foso que valdrá 10x, 100x o más, mucho más que, digamos, los chips de 7 nm frente a los de 3 nm, a menos que filtrar instantáneamente los secretos algorítmicos.⁷³

de IA algo-

En pocas palabras, creo que no proteger los secretos algorítmicos es probablemente la forma más probable en que China puede seguir siendo competitiva en la carrera de AGI. (Discutiré esto más adelante).

Es difícil exagerar lo mala que es la seguridad de los secretos algorítmicos en este momento. Entre los laboratorios, hay miles de personas con acceso a los secretos más importantes; básicamente, no hay verificación de antecedentes, silos, controles, seguridad de información básica, etc. Las cosas se almacenan en servicios SaaS fácilmente hackeables. La gente parlotea en las fiestas en San Francisco. Cualquiera, con todos los secretos en su se le podrían ofrecer 100 millones de dólares y ser reclutados para un laboratorio chino en cualquier punto.⁷⁴ Puedes... Basta con mirar a través de las ventanas de la oficina. Y

⁷³ A veces bromeo con que el laboratorio

Los avances rítmicos no se comparten con la comunidad investigadora estadounidense, pero

¡Se están compartiendo con la comunidad investigadora china!

⁷⁴ De hecho, he

y así sucesivamente. Hay muchos artículos y rumores que circulan por SF, que pretenden tener amplios detalles de varios avances algorítmicos de laboratorio.

que ByteDance envió un correo electrónico básicamente a todas las personas que estaban en el periódico Google Gemini para reclutarlos, ofreciéndoles L8 (un puesto muy alto, con un salario presumiblemente similar) y presentándoles diciendo que reportarían directamente al CTO en Estados Unidos de ByteDance.

La seguridad de los laboratorios de IA no es mucho mejor que la "seguridad de inicio aleatorio". Vender directamente los secretos de AGI al PCCh sería, al menos, más honesto.

... ¿Es esto lo que vemos en OpenAI o en cualquier otro laboratorio de IA estadounidense? No. De hecho, lo que vemos es lo contrario: el equivalente de seguridad del queso suizo. La penetración china en estos laboratorios sería trivialmente fácil utilizando cualquier número de métodos de espionaje industrial, como simplemente sobornar al equipo de limpieza para que coloque dongles USB en las computadoras portátiles. Mi propia suposición es que todos esos laboratorios de IA estadounidenses están completamente penetrados y que China está recibiendo descargas nocturnas de toda la investigación y el código de IA estadounidense AHORA MISMO...

Marc Andreessen

Si bien será difícil, creo que estos secretos son defendibles. Probablemente solo hay docenas de personas que realmente "necesitan Para conocer" los detalles clave de la implementación de un determinado avance algorítmico en un laboratorio determinado (incluso si un número mayor necesita conocer la idea básica de alto nivel), puede examinar, aislar y monitorear intensamente a estas personas, además de actualizar radicalmente la seguridad de la información.

Lo que requerirá la "superseguridad"

Hay muchas cosas que están al alcance de la mano en materia de seguridad en los laboratorios de IA. La mera adopción de las mejores prácticas de, por ejemplo, fondos de cobertura secretos o seguridad a nivel de datos de clientes de Google, nos pondría en una posición mucho mejor con respecto al espionaje económico "regular" del PCCh. De hecho, hay ejemplos notables de empresas del sector privado que lo han hecho notablemente bien en la preservación de secretos. Tomemos como ejemplo a las empresas de comercio cuantitativo (las Jane Streets del mundo). Varias personas me han dicho que en una hora de conversación podrían transmitir suficiente información a un competidor como para que el alfa de su empresa llegara a ~cero, similar a la cantidad de secretos algorítmicos clave de IA que podrían transmitirse en una conversación corta y, sin embargo, estas empresas logran mantener estos secretos y conservar su ventaja.

Si bien la mayoría de los principales laboratorios de IA de Estados Unidos se han negado a anteponer el interés nacional,

rechazando incluso las medidas básicas de seguridad.

Si tienen algún costo o requieren alguna priorización de seguridad, la recolección de esta fruta al alcance de la mano estaría dentro de sus capacidades.

Pero miremos un poco más allá. Una vez que China comience a comprender realmente la importancia de AGI, deberíamos esperar que toda la fuerza de sus esfuerzos de espionaje se haga realidad; pensemos en miles de millones de dólares invertidos, miles de empleados y medidas extremas (por ejemplo, equipos de ataque de operaciones especiales) dedicadas a infiltrarse en los esfuerzos de AGI estadounidenses. ¿Qué requerirá la seguridad de la AGI y la superinteligencia?

En resumen, *esto solo será posible con la ayuda del gobierno*. Microsoft, por ejemplo, es hackeado regularmente por actores estatales (por ejemplo, los hackers rusos **robaron recientemente los correos electrónicos de los ejecutivos de Microsoft**, así como **los correos electrónicos del gobierno que Microsoft aloja**). Un experto en seguridad de alto nivel que trabajaba en el campo estimó que incluso con un curso intensivo privado completo, China probablemente aún podría extraer los pesos de AGI si fuera su prioridad # 1: la única manera de llevar esta probabilidad a un solo dígito requeriría, más o menos, un proyecto gubernamental.

Si bien el gobierno no tiene un historial perfecto en materia de seguridad, es el único que tiene la infraestructura, el conocimiento y las competencias para proteger los secretos a nivel de defensa nacional. Cosas básicas como la autoridad para someter a los empleados a una intensa investigación; amenazar con penas de cárcel por filtrar secretos; seguridad física para centros de datos; y el vasto conocimiento de lugares como la NSA y las personas detrás de las autorizaciones de seguridad (las empresas privadas simplemente no tienen la experiencia en ataques de actores estatales).

No soy una de las personas detrás de las autorizaciones de seguridad, por lo que no puedo dar una explicación adecuada de lo que realmente requerirá la seguridad para AGI. El mejor recurso público sobre esto es el puerto electrónico RAND's **sobre seguridad de peso**. Para dar una idea de lo que realmente significará esta seguridad a prueba de actores estatales:

- Centros de datos totalmente aislados, con seguridad física a la par con la mayoría de las bases militares seguras (personal autorizado, fortificaciones físicas, equipo de respuesta in situ,

vigilancia exhaustiva

SITUATIONAL AWARENESS

y control de acceso extremo),

- Y no solo para los clústeres de entrenamiento, los clústeres de inferencia necesitan ¡La misma seguridad intensa!⁷⁵

⁷⁵ Es probable que las flotas de inferencia sean mucho más grandes que los clústeres de entrenamiento, y no hay será una presión abrumadora para medida que el espionaje chino y de otros países aumente, se actualizará rápidamente a un nivel mucho más intenso

- Nuevos avances técnicos en el cifrado confidencial de software duro y cómputo⁷⁶ y un escrutinio extremo de toda la cadena de suministro de hardware,
- Todo el personal investigador que trabaja en un SCIF (Servicio de Información Participada sensible, pronunciado "skiff", véase esta [visualización](#)),
- Verificación extrema del personal y autorizaciones de seguridad (incluidas pruebas periódicas de integridad de los empleados y similares), monitoreo constante y libertades sustancialmente reducidas para irse, y silos de información rígidos.
- Fuertes controles internos, por ejemplo, aprobación de múltiples teclas para ejecutar cualquier código.
- Limitaciones estrictas de las dependencias externas y satisfacción de los requisitos generales de las redes TS/SCI.
- Pruebas de penetración intensas en curso por parte de la NSA o similares.
- Y así sucesivamente. . .

Los gigantescos cúmulos de AGI se están mapeando, en este momento; El esfuerzo de seguridad correspondiente también debe serlo. Si estamos construyendo AGI en unos pocos años, tenemos muy poco tiempo.

Aun así, este inmenso esfuerzo no debe conducir al fatalismo. La gracia salvadora en materia de seguridad es que el PCCh probablemente aún no está completamente capacitado y, por lo tanto, aún no ha invertido en los esfuerzos más extremos. La seguridad de los laboratorios de IA estadounidenses "sólo" tiene que mantenerse a la vanguardia en comparación con la intensidad de los esfuerzos de espionaje chinos. Eso significa mejorar inmediatamente la seguridad para adelantarse al espionaje económico "más normal" (al que estamos lejos de resistirnos, pero las empresas privadas probablemente podrían serlo); y eso significa que en los próximos dos años, a

utilizar estos clústeres de inferencia para ejecutar investigadores de IA automatizados durante la explosión de inteligencia (y ejecutar miles de millones de superinteligencias de manera más amplia inmediatamente después). Por lo tanto, los pesos de AGI y superinteligencia también podrían extraerse de estos grupos. (Me preocupa que esto esté subestimado y que los clústeres de inferencia estén mucho menos protegidos).

⁷⁶ ¡Pero no puedes confiar solo en esto! El cifrado de hardware se **canaliza regularmente de forma lateral**. La defensa en profundidad es clave, por supuesto.

medidas en cooperación con el Gobierno.

Algunos ARGUMENTAN QUE las estrictas medidas de seguridad y su fricción asociada no valen la pena porque ralentizarían demasiado los laboratorios de IA estadounidenses. Pero creo que eso es un error:

- Esta es una tragedia del problema de los bienes comunes. Para los intereses comerciales de un laboratorio determinado, las medidas de seguridad que provocan una ralentización del 10 % pueden ser perjudiciales en competencia con otros laboratorios. Pero el interés nacional está claramente mejor servido si todos los laboratorios estuvieran dispuestos a aceptar la fricción adicional:
¿Puede la investigación de la IA estar muy por delante del progreso algorítmico chino y de otros países, y Estados Unidos que conserva un progreso algorítmico de velocidad del 90% como nuestra ventaja nacional es claramente mejor que retener el 0% como ventaja nacional (con todo robado instantáneamente)?
- Además, aumentar la seguridad ahora será el camino *menos* doloroso en términos de productividad de la investigación a largo plazo. Eventualmente, inevitablemente, aunque sólo sea en la cúspide de la superinteligencia, en la extraordinaria carrera armamentista que se avecina, el Gobierno de los Estados Unidos se dará cuenta de que la situación es insostenible y exigirá una represión de la seguridad. Será mucho más doloroso, y causará una ralentización mucho mayor, tener que implementar medidas de seguridad extremas, a prueba de actores estatales, desde un principio, en lugar de iterativamente.

Otros argumentan que incluso si nuestros secretos o pesos se filtran, aún lograremos salir *adelante* siendo más rápidos en otros aspectos (por lo que no debemos preocuparnos por las medidas de seguridad). Eso también es un error, o al menos correr demasiado riesgo:

- Como discutiré en un [artículo posterior](#), creo que el PCCh bien podría ser capaz de superar bruscamente a los EE. UU. (un clúster de 100 GW será mucho más fácil para ellos). En términos más generales, es posible que China no tenga la misma cautela que Estados Unidos (¡tanto una precaución razonable como una irrazonable!). Incluso si robar los algoritmos o los pesos

"sólo" los pone a la par con los EE. UU. en cuanto a modelos, eso podría ser suficiente para que ganen la carrera hacia la superinteligencia.

- Además, incluso si Estados Unidos sale adelante al final, la diferencia entre una ventaja de 1 a 2 años y de 1 a 2 meses será *realmente* importante para navegar por los peligros de la superinteligencia. Una ventaja de 1 a 2 años significa al menos un margen razonable para obtener la seguridad correcta y navegar por el período extremadamente volátil La explosión de inteligencia y la post-superinteligencia.⁷⁷ Un adicionales

Un mero adelanto de 1 a 2 meses significa una vertiginosa carrera armamentista internacional con presiones extremas, corriendo a través de la explosión de inteligencia, y sin espacio alguno para hacer las cosas bien en materia de seguridad. Es esa carrera existencial **en la que nos enfrentamos a los mayores riesgos de autodestrucción.**

- No te olvides de Rusia, Irán, Corea del Norte, etc. Sus capacidades de piratería no se quedan atrás. ¡En el curso actual, también estamos compartiendo libremente la superinteligencia con ellos! Sin una seguridad mucho mejor, estamos proliferando lo que será nuestra arma más poderosa para una plétora de Actores peligrosos, imprudentes e impredecibles.⁷⁸ energía nuclear

No vamos por buen camino

Cuando POR primera vez quedó claro para unos pocos que una bomba atómica era posible, el secretismo también fue quizás el tema más polémico. En 1939 y 1940, Leo Szilard se hizo conocido "en toda la comunidad física estadounidense como el principal apóstol del secreto en materia de fisión".⁷⁹ Pero fue rechazado por la mayoría; El secretismo era *bomba atómica, p. 509*

No era algo a lo que los científicos estuvieran acostumbrados, y que iba en contra de muchos de sus instintos básicos de ciencia abierta. Pero poco a poco se hizo evidente lo que había que hacer: el potencial militar de esta investigación era demasiado grande para que simplemente se compartiera libremente con los nazis. Y finalmente se impuso el secreto, justo a tiempo.

En el otoño de 1940, Fermi había terminado nuevas mediciones de absorción de carbono en grafito, lo que sugiere que el grafito era un moderador viable para una bomba. Szilard atacó a Fermi

⁷⁷ Por ejemplo, espacio para tomar 6 meses

durante la explosión de inteligencia para la investigación de alineación para asegurarse de que la superinteligencia no salga mal, Es hora de estabilizar la situación después de la invención de algunas nuevas armas de destrucción masiva dirigiendo estos sistemas para que se centren en aplicaciones defensivas, o simplemente tiempo para que los responsables humanos tomen las decisiones correctas dado el ritmo extraordinariamente rápido del cambio tecnológico con el advenimiento de la superinteligencia.

⁷⁸ Nos esforzamos mucho por evitar la

proliferación a estados canallas, incluso si todavía estaríamos "por delante" en tecnología nuclear en comparación con su arsenal más limitado, dado el caos proliferación.

⁷⁹ *La fabricación de la*



relató Szilard. Afortunadamente, las apelaciones posteriores finalmente tuvieron éxito, y Fermi se abstuvo a regañadientes de publicar sus resultados de grafito.

Figura 31: Leo Szilard.

Al mismo tiempo, el proyecto alemán se había reducido a dos posibles materiales moderadores: grafito y agua pesada. **A principios de 1941** en Heidelberg, Walther Bothe hizo una medición incorrecta de la sección transversal de absorción del grafito, y concluyó que el grafito absorbería demasiados neutrones para sostener una reacción en cadena. Dado que Fermi había mantenido su resultado en secreto, los alemanes no tenían las mediciones de Fermi para verificar y corregir el error. *Esto fue crucial: dejó que el proyecto alemán se dedicara al agua pesada, un camino equivocado decisivo que, en última instancia, condenó el esfuerzo alemán en materia de armas nucleares.*⁸⁰

⁸⁰ La fabricación de la bomba atómica, pág. 507

Si no fuera por esa apelación secreta de última hora, el proyecto de la bomba alemana podría haber sido un competidor mucho más formidable, y la historia podría haber resultado muy diferente.

Hay UNA DISONANCIA MENTAL real sobre la seguridad en los principales laboratorios de IA. Afirman a todo pulmón que están construyendo AGI en esta década. Enfatizan que el liderazgo estadounidense en materia de AGI será decisivo para la seguridad nacional de Estados Unidos. Según se informa, están planeando **la construcción de chips 7T** que solo tienen sentido si *realmente* crees en AGI. Y, de hecho, cuando mencionas la seguridad, asienten con la cabeza y reconocen "por supuesto, todos estaremos en un búnker" y sonrían.

Y, sin embargo, la realidad de la seguridad no podría estar más divorciada de eso. Cada vez que llega el momento de tomar decisiones difíciles para priorizar la seguridad, las actitudes de las startups y los intereses comerciales prevalecen sobre el interés nacional. El asesor de seguridad nacional tendría un colapso mental si entendiera el nivel de seguridad en los principales laboratorios de inteligencia artificial del país.

Hay secretos que se están desarrollando en este momento, que se pueden usar para cada carrera de entrenamiento en el futuro y serán los desbloques de claves para AGI, que están protegidos por la seguridad de una startup y valer cientos de miles de millones de dólares para el PCC.⁸¹ El real- a) en los próximos 12 a 24 meses, desarrollaremos los avances algorítmicos clave para AGI y los filtraremos rápidamente al

⁸¹ Eficiencias informáticas 100x+, cuando PCC, y b) ni siquiera estamos en camino de que nuestros

pesos estén seguros contra actores deshonestos como Corea del Norte, y mucho menos contra un

Se están construyendo grupos por valor de 10 o 100 millones de dólares.

un esfuerzo total por parte de China, para el momento en que construyamos AGI. "Buena seguridad para una startup" simplemente no es lo suficientemente buena, y tenemos muy poco tiempo antes de que el daño atroz a la seguridad nacional de los Estados Unidos se vuelva irreversible.

Estamos desarrollando el arma más poderosa que la humanidad haya creado jamás. Los secretos algorítmicos que estamos desarrollando, en este momento, son literalmente los secretos de defensa nacional más importantes de la nación, los secretos que estarán en la base del predominio económico y militar de EE. UU. y sus aliados para fines de la década, los secretos que determinarán si tenemos el liderazgo necesario para lograr la seguridad correcta de la IA. los secretos que determinarán el resultado de la Tercera Guerra Mundial, los secretos que determinarán el futuro del mundo libre. Y, sin embargo, la seguridad de los laboratorios de IA es probablemente peor que la de un contratista de defensa al azar que fabrica pernos.

Es una locura.

Básicamente, nada de lo que hagamos, en la competencia nacional y en la seguridad de la IA, importará si no solucionamos esto pronto.



Figura 32: Unavalla publicitaria en las instalaciones de enriquecimiento de uranio de Oak Ridge, Tennessee, 1943.

IIIc. Superalineación

Controlar de forma fiable sistemas de IA mucho más inteligentes que nosotros es un problema técnico sin resolver. Y aunque es un problema solucionable, las cosas podrían descarrilarse muy fácilmente durante un Rápida explosión de inteligencia. Manejar esto será extremadamente tenso; El fracaso podría ser fácilmente catastrófico.

El viejo hechicero
¡Finalmente se ha ido!
Ahora los espíritus que él
controla obedecerán mis
órdenes.

...

Yo también haré
maravillas.

...

¡Señor, estoy en una
situación desesperada!

*Los espíritus que
convoqué, no puedo
deshacerme de ellos.*

Johann WOLFGANG VON GOETHE

"El aprendiz de brujo"

A estas ALTURAS , probablemente hayas oído hablar de los agoreros de la IA. Es posible que te hayan intrigado sus argumentos, o que los hayas descartado de plano. Eres reacio a leer otra meditación pesimista.

No soy un agorero.⁸² La superinteligencia desalineada probablemente no sea el mayor riesgo de la IA.⁸³ Pero pasé el año pasado trabajando en la investigación técnica sobre la alineación de los sistemas de IA como mi trabajo diario en OpenAI, trabajando con Ilya y la **Superalineación**

⁸² ¡Como lo certifié no muy cortésmente el agorero en jefe! Por lo menos, me llamaría a mí mismo un fuerte optimista de que este problema tiene solución. He gastado considerables energías en debatir la Los pesimistas de la IA y abogaron firmemente contra políticas como la pausa de la IA.

⁸³ Lo que más me preocupa es que las cosas estén totalmente locas en torno a la superinteligencia, incluidas cosas como las nuevas armas de destrucción masiva, las guerras destructivas y las

incógnitas desconocidas. Por otra parte, creo que el arco de la historia nos aconseja no subestimar el autoritarismo, y la superinteligencia podría permitir que los autoritarios dominen durante miles de millones de años.

equipo. Hay un problema técnico muy real: nuestras técnicas de alineación actuales (métodos para garantizar que podamos controlar, dirigir y confiar de manera confiable en los sistemas de IA) no se escalarán a sistemas de IA sobrehumanos. Lo que quiero hacer es explicar lo que veo como el "predeterminado" planear cómo vamos a salir adelante,⁸⁴ y por qué soy optimista. Si bien no hay suficientes personas en la pelota (¡deberíamos tener esfuerzos mucho más ambiciosos para resolver este problema!), en general, hemos tenido suerte con la forma en que el aprendizaje profundo se ha sacudido, hay una gran cantidad de fruta empírica al alcance de la mano que nos llevará parte del camino, y tendremos la ventaja de millones de investigadores de IA autoacoplados para llevarnos el resto del camino.

Pero también quiero decirte por qué estoy preocupado. Sobre todo, asegurar que la alineación no salga mal requerirá una competencia extrema en la gestión de la explosión de inteligencia. Si hacemos una transición rápida de la AGI a la superinteligencia, nos enfrentaremos a una situación *en la que, en menos de un año*, pasaremos de sistemas reconocibles a nivel humano para los que los descendientes de las técnicas de alineación actuales funcionarán en su mayoría bien, a sistemas mucho más extraños y vastamente sobrehumanos que plantean un problema de alineación técnica cualitativamente diferente y fundamentalmente novedoso; al mismo tiempo, pasar de sistemas en los que el fracaso es de bajo riesgo a sistemas extremadamente poderosos en los que el fracaso puede ser catastrófico; Todo mientras la mayor parte del mundo probablemente se está volviendo un poco loco. Me pone bastante nerviosa.

Para cuando termine la década, tendremos miles de millones de agentes de IA inmensamente superiores corriendo por ahí. Estos agentes sobrehumanos de IA serán capaces de realizar actividades extremadamente complejas y creativas; No tendremos ninguna esperanza de seguir adelante. Seremos como los estudiantes de primer grado tratando de supervisar con múltiples doctorados.

En esencia, nos enfrentamos a un problema de transferencia de *confianza*. Al final de la explosión de inteligencia, no tendremos ninguna esperanza de comprender lo que están haciendo nuestros mil millones de superinteligencias (excepto en la medida en que decidan explicárnoslo, como lo harían con un niño). Y todavía no tenemos la capacidad técnica para garantizar de manera confiable incluso las

⁸⁴ Como dice Tyler Cowen, confundir restricciones secundarias básicas para estos sistemas, como "no mientas" o "sigue la ley" o "no intentes filtrar tu servidor". El refuerzo de la retroalimentación humana (RLHF) funciona muy bien para agregar tales restricciones laterales para la corriente

¡A través de está subestimado!

pero RLHF depende de que los humanos sean capaces de comprender y supervisar el comportamiento de la IA, que fundamentalmente no se escalará a sistemas sobrehumanos.

En pocas palabras, sin un esfuerzo muy concertado, no podremos garantizar que la superinteligencia no se vuelva rebelde (y esto es reconocido por muchos líderes en el campo). Sí, todo puede estar bien por defecto. Pero aún no lo sabemos. Especialmente una vez que los futuros sistemas de IA no solo se entrenen con el aprendizaje por imitación, sino también con RL (aprendizaje por refuerzo) a gran escala y a largo plazo, adquirirán comportamientos propios e impredecibles, moldeados por un proceso de prueba y error (por ejemplo, pueden aprender a mentir o buscar poder, ¡simplemente porque estas son estrategias exitosas en el mundo real!).

Lo que está en juego será lo suficientemente alto como para que esperar lo mejor simplemente no sea una respuesta lo suficientemente buena en cuanto a la alineación.

El problema

El problema de la superalineación

Hemos sido capaces de desarrollar un método muy exitoso para alinear (es decir, dirigir/controlar) *los sistemas de IA actuales* (¡sistemas de IA más tontos que nosotros!): el Aprendizaje por Refuerzo a partir de la Retroalimentación Humana (RLHF). La idea detrás de RLHF es simple: el sistema de IA prueba cosas, los humanos califican si su comportamiento fue bueno o malo, y luego refuerzan los buenos comportamientos y penalizan los malos comportamientos. De esa manera, aprende a seguir las preferencias humanas.

De hecho, RLHF ha sido la clave detrás del éxito de ChatGPT y otros. Los modelos base de 85 tenían un montón de inteligencia en bruto, pero no los aplicaban de una manera útil por defecto; por lo general, solo respondían con un lío confuso que se asemejaba a un texto aleatorio de Internet. A través de RLHF, podemos dirigir su comportamiento, inculcando conceptos básicos importantes como el seguimiento de instrucciones y la ayuda. RLHF también nos permite hornear en barandillas de seguridad: por ejemplo, si un usuario me pide instrucciones

sobre armas biológicas, el modelo probablemente debería negarse.⁸⁶

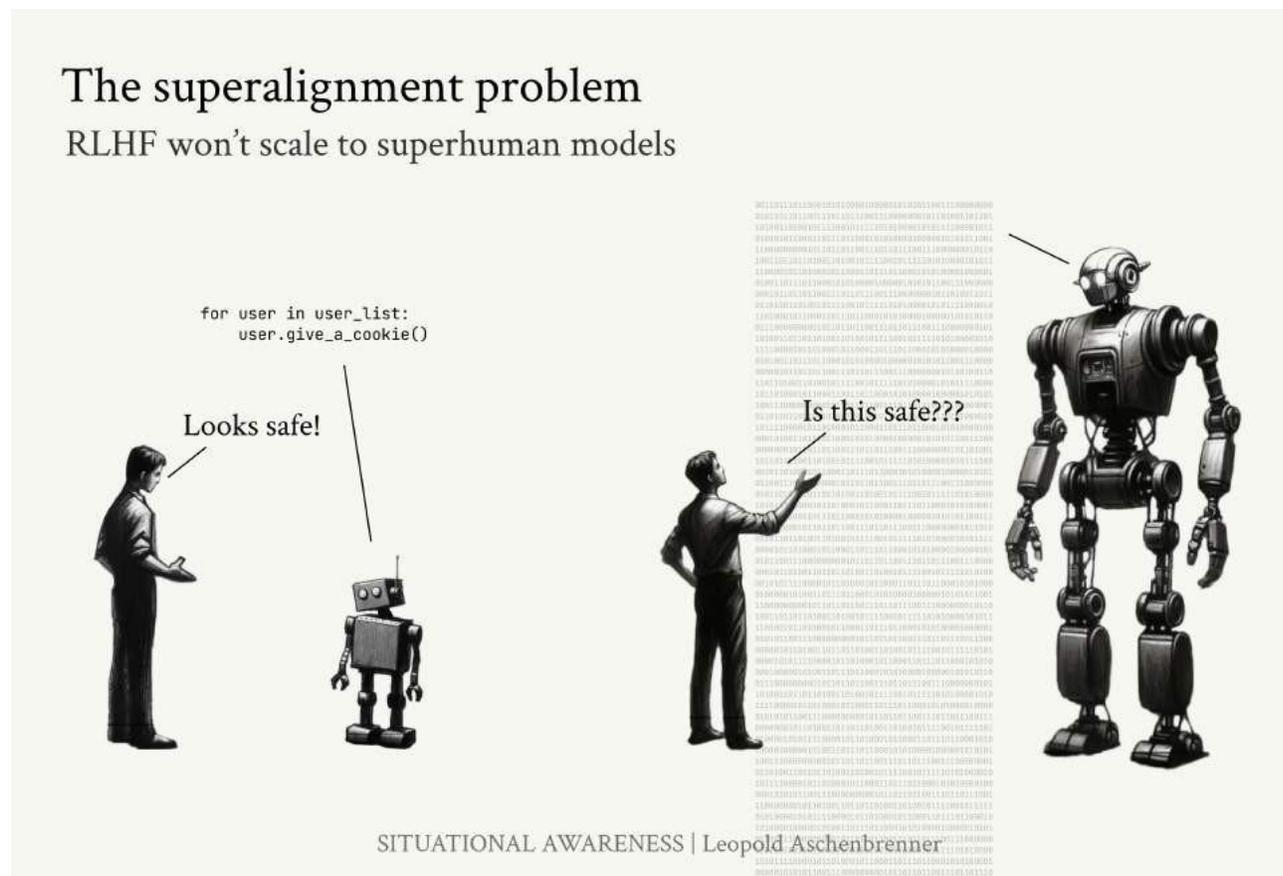
⁸⁵ Irónicamente, los chicos de seguridad hicieron el mayor avance para permitir el éxito comercial de la IA al inventar RLHF. Los modelos base tenían mucha inteligencia en bruto, pero eran inmanejables y, por lo tanto, inutilizables para la mayoría de las aplicaciones.

⁸⁶ Esto pone de relieve una distinción importante: la *capacidad técnica* de alinear (dirigir/controlar) un modelo es independiente de una cuestión de valores *sobre a qué alinearse*.

Ha habido muchas controversias políticas sobre esta última cuestión. Y aunque estoy de acuerdo con la oposición a algunas de las excrecencias aquí, eso no debería distraer la atención del problema técnico básico. Sí, las técnicas de alineación pueden ser mal utilizadas, pero necesitaremos mejores técnicas de alineación para garantizar incluso las restricciones laterales básicas para los modelos futuros, como seguir las instrucciones o seguir la ley. Véase también "[La alineación de la IA es distinta de sus aplicaciones a corto plazo](#)".

**El problema técnico central de la superalineación es simple:
¿cómo controlamos los sistemas de IA (mucho) más
inteligentes que nosotros?**

Como era de esperar, RLHF se desmoronará a medida que los sistemas de IA se vuelvan más inteligentes, y nos enfrentaremos a desafíos técnicos fundamentalmente nuevos y cualitativamente diferentes. Imaginemos, por ejemplo, un sistema de IA sobrehumano que genera un millón de líneas de código en un nuevo lenguaje de programación que ha inventado. Si le preguntara a un evaluador humano en un procedimiento de RLHF, "¿este código contiene alguna puerta trasera de seguridad?", simplemente no lo sabría. No podrían calificar el resultado como bueno o malo, seguro o inseguro, por lo que no sería capaz de reforzar los buenos comportamientos y penalizar los malos comportamientos con RLHF.



Incluso ahora, los laboratorios de IA ya **tienen que pagar a ingenieros de software expertos** para que otorguen calificaciones RLHF para el código de ChatGPT: ¡el código que pueden generar los modelos actuales ya está bastante avanzado! El pago de los etiquetadores humanos ha pasado de unos pocos dólares para las etiquetadoras MTurk a

Figura 33: La alineación de los sistemas de IA a través de la supervisión humana (como en RLHF) no se escalará a la superinteligencia.

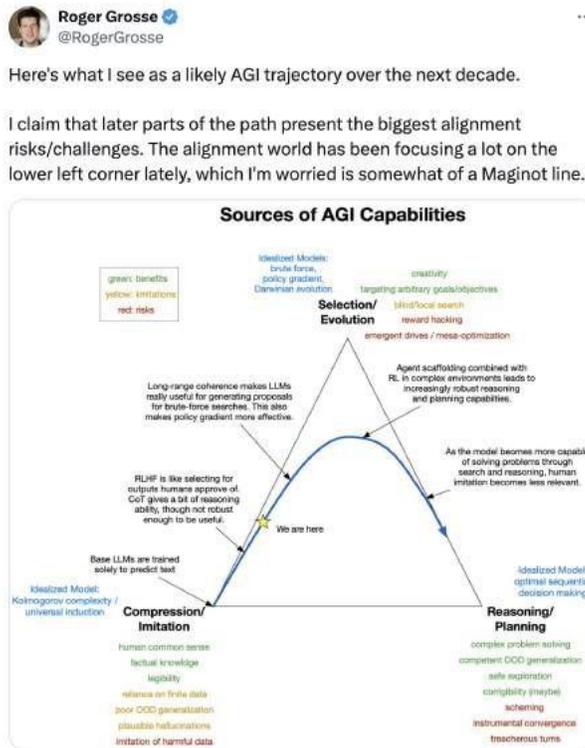
~\$100/hora para preguntas de GPQA87 en los últimos años. En un futuro (cercano), incluso los mejores expertos humanos que dediquen mucho tiempo no serán lo suficientemente buenos. Estamos empezando a encontrarnos con las primeras versiones del problema de la superalineación en el mundo real, y muy pronto esto será un problema importante, incluso para el despliegue práctico de los sistemas de próxima generación. Está claro que necesitaremos un sucesor de RLHF que se adapte a las capacidades de IA mejor que a nivel humano, donde la supervisión humana se rompe. En cierto sentido, el objetivo de los esfuerzos de investigación de superalineación es repetir la historia de éxito de RLHF: hacer las apuestas de investigación básica que serán necesarias para dirigir y desplegar sistemas de IA dentro de un par de años.

⁸⁷ "Estimamos un promedio por hora pago de aproximadamente \$95 por hora", p.3 del documento de GPQA.

Cómo se ve el fracaso

Con demasiada frecuencia, la gente se imagina un "chatbot GPT-6", informando a sus intuiciones de que seguramente no estarían peligrosamente desalineados. Como se discutió anteriormente en esta serie, la trayectoria de "desaceleración" apunta a agentes, entrenados con RL, en un futuro cercano. Creo que el gráfico de Roger lo hace bien.

Figura 34: Roger Grossesobre la evolución del problema de la alineación a medida que los sistemas se vuelven más avanzados.



Una forma de pensar en lo que estamos tratando de lograr con la alineación, desde la perspectiva de la seguridad, es agregar *restricciones laterales*. Consideremos un futuro y poderoso "modelo base" que, en una segunda etapa de entrenamiento, entrenamos con RL de largo horizonte para ejecutar un negocio.

ness y ganar dinero⁸⁸ (como ejemplo simplificado):
de IA

- Por defecto, bien puede aprender a mentir, a cometer fraude, a engañar, a hackear, a buscar poder, etc., simplemente porque estas pueden ser estrategias exitosas para ganar dinero en la vida real
¡mundo!⁸⁹
explorar-
- Lo que queremos es añadir restricciones secundarias: no mentir, no infringir la ley, etc.
- Pero aquí volvemos a la cuestión fundamental de la alineación de los sistemas sobrehumanos: no seremos capaces de entender lo que están haciendo, y por lo tanto no seremos capaces de notar y penalizar el mal comportamiento con RLHF.⁹⁰
inferencia

Si no podemos agregar estas restricciones secundarias, no está claro qué sucederá. *Tal vez* tengamos suerte y las cosas sean benignas por defecto (por ejemplo, tal vez podamos llegar bastante lejos sin que los sistemas de IA tengan objetivos a largo plazo, o los comportamientos indeseables serán menores). Pero también es totalmente plausible que aprendan comportamientos indeseables mucho más serios: aprenderán a mentir, aprenderán a buscar el poder, aprenderán a comportarse bien cuando los humanos están mirando y a perseguir estrategias más nefastas cuando no estamos mirando, y así sucesivamente.

El hecho de que el problema de la superalineación no se resuelva significa que simplemente no tendremos la capacidad de garantizar ni siquiera estas restricciones secundarias básicas para estos sistemas de superinteligencia, como "*¿seguirán mis instrucciones de manera confiable?*" o "*¿responderán honestamente a mis preguntas?*" o "*¿No engañarán a los humanos?*". A menudo, las personas asocian la alineación con algunas preguntas complicadas sobre los valores humanos, o saltan a las controversias políticas, pero decidir qué comportamientos y valores inculcar en el modelo, si bien es importante,

⁸⁸ Muy simplificado, piense en un sistema

intentando, a través de prueba y error, maximizar dinero durante un período de un año, y el El modelo de IA final entrenado es el resultado de un proceso de selección que selecciona los sistemas de IA que tuvieron más éxito en maximizar el dinero.

⁸⁹ Lo que RL está haciendo es simplemente

estrategias para tener éxito en el objetivo. Si una estrategia funciona, es reforzado en el modelo. Por lo tanto, si la mentira, el fraude, la búsqueda de poder, etc. (o los patrones de pensamiento que podrían conducir a este tipo de comportamientos en al menos algunas situaciones) funcionan, estos también se reforzarán en el modelo.

⁹⁰ (o modelos de monitoreo de tiempo de

entrenado con supervisión humana)

es un problema aparte. El problema principal es que para lo que sea que quieras inculcar el modelo (¡incluyendo garantizar cosas muy básicas, como "seguir la ley"!) todavía no sabemos cómo hacerlo para los sistemas de IA muy poderosos que estamos construyendo muy bien.

pronto.

Una vez más, las consecuencias de esto no están del todo claras. Lo que está claro es que la superinteligencia tendrá vastas capacidades, por lo que el mal comportamiento podría ser catastrófico con bastante facilidad. Es más, espero que dentro de un pequeño número de años, estos sistemas de IA se integren en muchos sistemas críticos, incluidos los sistemas militares (de lo contrario, significaría la dominación total de los adversarios). Parece una locura, pero ¿recuerdas cuando todo el mundo decía que no conectaríamos la IA a Internet? Lo mismo ocurrirá con cosas como "¡nos aseguraremos de que un humano esté siempre informado!"—como se dice hoy en día.

Los fallos de alineación pueden parecer incidentes aislados, por ejemplo, un agente autónomo que comete fraude, una instancia modelo que se autoexfiltra, un investigador automatizado que falsifica una experiencia. resultado mental, o un enjambre de drones que sobrepasa las reglas de compromiso. Pero los fracasos también podrían ser a una escala mucho mayor o más sistemáticos: en el extremo, los fracasos podrían parecerse más a una rebelión de robots. Habremos convocado a una inteligencia bastante extraña, mucho más inteligente que nosotros, cuya arquitectura y proceso de entrenamiento ni siquiera fueron diseñados por nosotros, sino por una generación anterior de sistemas de IA superinteligentes, en la que ni siquiera podemos empezar a entender lo que están haciendo, estará dirigiendo nuestro ejército y sus objetivos habrán sido aprendidos por un proceso de selección natural.

A menos que resolvamos la alineación, a menos que descubramos cómo inculcar esas restricciones secundarias, no hay ninguna razón particular para esperar que esta pequeña civilización de superinteligencias continúe obedeciendo órdenes humanas a largo plazo. Parece totalmente dentro del ámbito de las posibilidades que en algún momento simplemente se inventen para eliminar a los humanos, ya sea repentina o gradualmente.

La explosión de inteligencia hace que todo esto sea increíblemente tenso

Soy optimista en cuanto a que la superalineación es un problema técnico que se puede resolver. Al igual que desarrollamos RLHF, podemos desarrollar el sucesor de RLHF para sistemas sobrehumanos y hacer el ciencia que nos da una gran confianza en nuestros métodos.

Si las cosas siguen progresando de forma iterativa, si insistimos en una seguridad rigurosa

pruebas y demás, todo debería ser factible (y discutiré mi mejor conjetura actual de cómo nos las arreglaremos más en un momento).

Lo que hace que esto sea increíblemente espeluznante es la posibilidad de una explosión de inteligencia: que podríamos hacer la transición de sistemas más o menos humanos a *sistemas sobrehumanos extremadamente rápido, tal vez en menos de un año*.

Alignment during the intelligence explosion

	AGI	Superintelligence
Required alignment technique	RLHF++	Novel, qualitatively different technical solutions
Failures	Low-stakes	Catastrophic
Architecture and algorithms	Familiar, descendants of current systems, fairly benign safety properties	Alien. Designed by previous-generation super-smart AI system
Backdrop	World is normal	World is going crazy, extraordinary pressures
Epistemic state	We can understand what the systems are doing, how they work, and whether they're aligned.	We have no ability to understand what's going on, how to tell if systems are still aligned and benign, what the systems are doing, and we are entirely reliant on trusting the AI systems.

Transition in <1 year?
 with very little time to get decisions right

SITUATIONAL AWARENESS | Leopold Aschenbrenner

- Pasaremos extremadamente rápido de sistemas en los que RLHF funciona bien a sistemas en los que se descompondrá por completo. Esto nos deja muy poco tiempo para descubrir y

Figura 35: La explosión de inteligencia hace que la superalineación sea increíblemente espeluznante.

abordar las formas en que nuestros métodos actuales fracasarán.

- Al mismo tiempo, pasaremos extremadamente rápido de sistemas en los que los fracasos son bastante poco arriesgados (ChatGPT dijo una mala palabra, ¿y qué?) a sistemas extremadamente altos (ups, la superinteligencia se autofiltró de nuestro clúster, ahora está hackeando a los militares). En lugar de encontrarnos iterativamente con fallos de seguridad cada vez más peligrosos en la naturaleza, los primeros fallos de seguridad notables que encontramos podrían ser catastróficos.
- La superinteligencia que obtengamos al final será *inmensamente* sobrehumana. Dependemos enteramente de confiar en estos sistemas, y confiar en lo que nos dicen que está pasando, ya que ya no tendremos capacidad propia para penetrar en lo que están haciendo exactamente.
- La superinteligencia que obtenemos al final podría ser bastante *alienígena*. Habremos pasado por una década o más de avances en ML durante la explosión de inteligencia, lo que significa que las arquitecturas y los algoritmos de entrenamiento serán totalmente diferentes (con propiedades de seguridad potencialmente mucho más riesgosas).
 - Un ejemplo que es muy destacado para mí: bien podemos arrancar nuestro camino hacia el nivel humano o algo sobrehumano de AGI con sistemas que razonan a través de cadenas de pensamientos, es decir, a través de fichas en inglés. Esto es *extraordinariamente* útil, porque significa que los modelos "piensan en voz alta" y nos permiten captar el comportamiento de los delincuentes (por ejemplo, si está conspirando en nuestra contra). Pero *seguramente* hacer que los sistemas de IA piensen en tokens no es el medio más eficiente para hacerlo, *seguramente* hay algo mucho mejor que hace todo este pensamiento a través de estados internos, por lo que es casi seguro que el modelo al final de la explosión de inteligencia *no* pensará en voz alta, es decir, tendrá un razonamiento completamente ininterpretable.
- Este será un período increíblemente volátil, potencialmente con el telón de fondo de una carrera armamentista internacional, toneladas de presión
Para ir más rápido, nuevas y salvajes capacidades avanzan cada semana básicamente sin tiempo humano para tomar

buenas decisiones, y así sucesivamente. Nos enfrentaremos a toneladas de datos ambiguos y decisiones de alto riesgo.

- Piensa: "Descubrimos que el sistema de IA hacía algunas cosas traviesas en una prueba, pero ajustamos un poco nuestro procedimiento para resolverlo. Nuestros investigadores de IA automatizada nos dicen que las métricas de alineación se ven bien, pero realmente no entendemos lo que está sucediendo y no confiamos plenamente en ellas, y no tenemos ningún conocimiento científico sólido que nos haga confiar en que esto seguirá siendo válido durante otro par de OOM. Entonces, ¿probablemente estaremos bien? Además, China acaba de robarnos nuestros pesos y está lanzando su propia explosión de inteligencia, nos pisa los talones".

Realmente parece que esto podría descarrilarse. Para ser honesto, suena aterrador.

Sí, contaremos con sistemas de IA que nos ayudarán. Al igual que automatizan la investigación de capacidades, podemos usarlos para automatizar la investigación de alineación. Esa será la clave, como discutiré a continuación. Pero, ¿se puede confiar en los sistemas de IA? En primer lugar, no estabas seguro de si estaban alineados, ¿están siendo realmente honestos contigo sobre sus afirmaciones sobre la ciencia de la alineación? ¿Podrá la investigación de alineación automática mantenerse al día con la investigación automatizada de capacidades (por ejemplo, porque la automatización de la alineación es más difícil, por ejemplo, porque hay métricas menos claras en las que podemos confiar en comparación con la mejora de las capacidades del modelo, o hay mucha presión para avanzar a toda velocidad en el progreso de las capacidades debido a la carrera internacional)? Y la IA no podrá sustituir por completo a los tomadores de decisiones aún humanos que toman buenas decisiones en esta situación increíblemente de alto riesgo.

El plan por defecto: cómo podemos salir del paso

Creo QUE podemos cosechar ganancias a través de una serie de apuestas empíricas, que describiré a continuación, para alinear sistemas algo sobrehumanos. Entonces, si estamos seguros de que podemos confiar en estos sistemas, tendremos que usar estos sistemas algo sobrehumanos para automatizar la investigación de alineación, junto con la automatización de la investigación de IA en general, durante la explosión de inteligencia, para descubrir cómo resolver la alineación para ir por el resto del camino.

Alinear modelos algo sobrehumanos

Alinear los sistemas a nivel humano no será suficiente. Incluso los primeros sistemas que puedan realizar investigaciones automatizadas de IA, es decir, iniciar la explosión de la inteligencia, probablemente ya serán sustancialmente sobrehumanos en muchos dominios. Esto se debe a que es probable que las capacidades de la IA sean algo puntiagudas: para cuando la AGI esté al nivel humano en lo que sea que un investigador/ingeniero de IA humano sea peor, será sobrehumana en muchas otras cosas. Por ejemplo, tal vez la capacidad de los sistemas de IA para coordinarse y planificar de manera efectiva se quede atrás, lo que significa que para cuando la explosión de inteligencia esté en pleno apogeo, probablemente ya serán codificadores sobrehumanos, que enviarán solicitudes de extracción de millones de líneas en nuevos lenguajes de programación que idearon, y serán sobrehumanos en matemáticas y ML.

Estos sistemas de explosión de inteligencia temprana comenzarán a ser cuantitativa y cualitativamente sobrehumanos, al menos en muchos dominios. Pero se parecerán mucho más a los sistemas que tenemos hoy en día en términos de arquitectura, y la brecha de inteligencia que necesitamos cubrir es mucho más manejable. (Tal vez si los humanos que intentan alinear la verdadera superinteligencia son como un niño de primer grado tratando de supervisar a un graduado de doctorado, esto es más como un estudiante de secundaria inteligente tratando de supervisar a un graduado de doctorado).

En términos más generales, cuanto más podamos desarrollar buena ciencia ahora, más estaremos en condiciones de verificar que las cosas no se están descarrilando durante la explosión de inteligencia. Incluso tener buenas métricas en las que podamos confiar para la superalineación es sorprendentemente difícil, pero sin métricas confiables durante la explosión de inteligencia, no sabremos si presionar es seguro o no.

Estas son algunas de las principales apuestas de investigación que veo para cruzar la brecha entre los sistemas a nivel humano y algo sobrehumanos.

LA EVALUACIÓN ES MÁS FÁCIL QUE LA GENERACIÓN. OBTENEMOS PARTE DEL camino "gratis", porque es más fácil para nosotros evaluar los resultados (especialmente para los malos comportamientos atroces) que generarlos nosotros mismos. Por ejemplo, me lleva

meses o años de arduo trabajo escribir un artículo, pero solo un par de horas para saber si un artículo que alguien ha escrito es bueno (aunque tal vez más

para detectar fraudes). Tendremos equipos de humanos expertos que pasarán mucho tiempo evaluando cada ejemplo de RLHF, y podrán "bajar" el pulgar hacia abajo de muchos malos comportamientos, incluso si el sistema de IA es algo más inteligente que ellos. Dicho esto, esto solo nos llevará hasta cierto punto (GPT-2 o incluso GPT-3 no pudieron detectar el nefasto GPT-4 de manera confiable, ¡a pesar de que la evaluación es más fácil que la generación!)

Supervisión ESCALABLE. Podemos utilizar los asistentes de IA para ayudar a los humanos a supervisar otros sistemas de IA, ya que el equipo humano-IA puede extender la supervisión más allá de lo que el humano podría hacerlo por sí solo. Por ejemplo, imaginemos que un sistema de IA genera 1 millón de líneas de código complejo. Es más fácil para un humano comprobar que una línea de código determinada tiene un error si un asistente de IA ha señalado que (por ejemplo) la línea 394.894 parece sospechosa, que para un humano encontrar ese mismo error desde cero. Un modelo entrenado para criticar el código escrito por otro modelo podría, por lo tanto, ayudar a los humanos a supervisar un sistema con habilidades de codificación estrechamente sobrehumanas.

Se han propuesto varias estrategias de supervisión escalables, entre ellas **el debate, la creación de mercado, el modelado de recompensas recursivas y los juegos de verificación de probadores**, así como versiones simplificadas de esas ideas, como **las críticas**. Los modelos son ahora lo suficientemente fuertes como para que sea posible **probar empíricamente estas ideas, lo que hace que se avance directamente en la supervisión escalable**.

Espero que esto ayude mucho para la parte "cuantitativamente" sobrehumana del problema, como el ejemplo de un millón de líneas de código antes mencionado. Pero soy menos optimista en cuanto a que las técnicas de supervisión escalables puedan ayudar realmente a la parte "cualitativamente" superhumana del problema: imagina que el modelo inventa la física cuántica cuando sólo entiendes la física newtoniana.

GENERALIZACIÓN. INCLUSO con una supervisión escalable, no podremos supervisar los sistemas de IA en problemas realmente difíciles, problemas que escapan a la comprensión humana. Sin embargo, podemos estudiar: ¿cómo generalizarán los sistemas de IA desde la supervisión

humana en problemas fáciles (que entendemos y podemos supervisar) hasta comportarse en los problemas difíciles (que no podemos entender y ya no podemos supervisar)? Por ejemplo, tal vez supervisando
Un modelo, para ser honesto en casos simples, generaliza benignamente a

El modelo es honesto en general, incluso en los casos en los que está haciendo cosas extremadamente complicadas que no entendemos.

Hay muchas razones para ser optimistas aquí: parte de la magia del aprendizaje profundo es que a menudo se generaliza de manera benigna (por ejemplo, RLHF con solo etiquetas en ejemplos en inglés también tiende a producir un buen comportamiento cuando se habla francés o español, aunque no formara parte de la formación). Soy bastante optimista de que habrá métodos bastante simples que ayuden a empujar la generalización de los modelos a nuestro favor, y que podamos desarrollar una sólida comprensión científica que nos ayude a predecir cuándo funcionará la generalización y cuándo fallará. En mayor medida que para la supervisión escalable, la esperanza es que esto ayude con la alineación incluso en el caso "cualitativamente" sobrehumano.

Aquí hay otra forma de pensar en esto: si un modelo sobrehumano se está comportando mal, digamos violando la ley, intuitivamente *el modelo ya debería saber* que está violando la ley. Además, "¿está esto violando la ley?" es probablemente un concepto bastante natural para el modelo, y será prominente en el espacio de representación del modelo. La pregunta entonces es: ¿podemos "convocar" este concepto del modelo con una supervisión débil?

Soy particularmente partidario de esta dirección (y tal vez sesgado), porque ayudé a introducir esto con un trabajo reciente con algunos colegas de OpenAI. En particular, se estudió una *¿Puede un modelo pequeño alinear a un modelo más grande (más inteligente)?* Descubrimos que la generalización en realidad te hace cruzar un poco (pero ciertamente no toda) de la brecha de inteligencia entre el supervisor y el supervisor, y que en entornos simples hay mucho que se puede hacer para mejorarla.

INTERPRETABILIDAD. Una forma intuitivamente atractiva de verificar y confiar en que nuestros sistemas de IA están alineados es si pudiéramos entender lo que están pensando. Por ejemplo, si nos preocupa que los sistemas de IA nos engañen o conspiren contra nosotros, el acceso a su razonamiento interno debería ayudarnos a detectarlo. Por defecto, los sistemas modernos de IA son cajas

negras inescrutables.

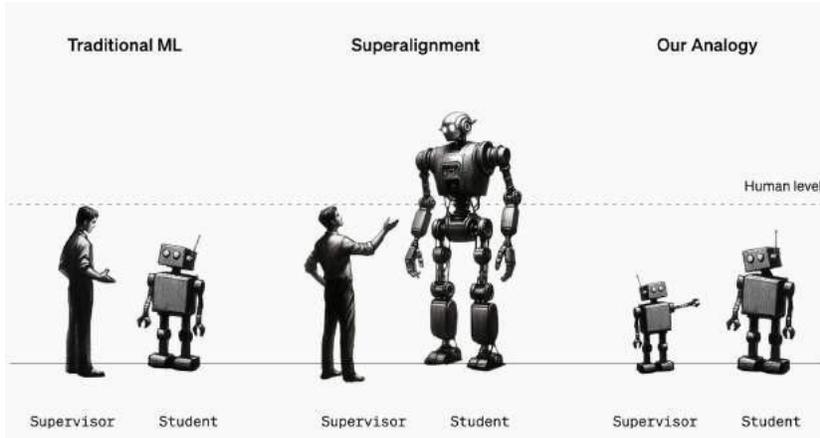


Figura 36: Una analogía simple para estudiar la superalineación: en lugar de un humano supervisando un modelo sobrehumano, podemos estudiar un modelo pequeño supervisando un modelo grande. Por ejemplo, ¿podemos alinear GPT-4 solo con la supervisión de GPT-2?

¿Dará lugar a que GPT-4 generalice adecuadamente "lo que GPT-2 significó"?

De la generalización débil a la fuerte.

Sin embargo, parece que deberíamos ser capaces de hacer una "neurociencia digital" asombrosa, después de todo, tenemos un acceso perfecto a modelos internos.

Aquí hay algunos enfoques diferentes, que van desde "más ambicioso y 'genial' pero será muy difícil" hasta "cosas más hackeadas que son más fáciles y podrían funcionar".

Interpretabilidad mecanicista. Intente aplicar ingeniería inversa a las grandes redes neuronales desde cero, desenredar por completo las matrices inescrutables, por así decirlo.

El equipo de Chris Olah en Anthropic ha hecho **gran parte del trabajo pionero-neering** en esto, comenzando por la comprensión de mecanismos simples en modelos muy pequeños. Recientemente se está produciendo un **progreso increíblemente emocionante**, y estoy encantado con el nivel general de actividad en este espacio.

Dicho esto, me preocupa que la ingeniería inversa completa de los sistemas de IA sobrehumana sea un problema intratable, similar a, por ejemplo, "ingeniería inversa completa del cerebro humano", y pondría este trabajo principalmente en el cubo de "ambicioso moonshot para la seguridad de la IA" en lugar de "plan predeterminado para salir del paso".⁹¹

(Los 200 problemas abiertos de Neel Nanda en la **interpretabilidad mecanicista** también dan una idea de este tipo de investigación).

⁹¹ Por otra parte, incluso los recientes avances en la interpretabilidad mecanicista "Desenredar las características" de los modelos con autocodificadores dispersos no resuelve por sí solo el problema de cómo lidiar con modelos sobrehumanos. Por un lado, el modelo podría simplemente "pensar" en conceptos sobrehumanos que no entiendes. Además, ¿cómo saber qué característica es la que ¿querer? Todavía no tienes etiquetas de verdad fundamental. Por ejemplo, puede haber toneladas de características diferentes que se parezcan a la "característica de la verdad" para usted, una de las cuales es "lo que el modelo realmente sabe" y los otros son "lo que pensaría el humano xyz" o "lo que los evaluadores humanos quieren que piense", etc.

Los escasos autocodificadores no

serán suficientes por sí solos, pero serán una herramienta -¡una herramienta increíblemente útil!- que en última instancia tiene que sacar provecho para ayudar con cosas como la ciencia de la generalización.

Interpretabilidad "de arriba hacia abajo". Si la interpretabilidad mecanicista intenta aplicar ingeniería inversa a las redes neuronales "de abajo hacia arriba", otros trabajos adoptan un enfoque más específico, "de arriba hacia abajo", tratando de localizar la información en un modelo sin una comprensión completa de cómo se procesa.

Por ejemplo, podríamos intentar construir un "detector de mentiras de IA" identificando las partes de la red neuronal que se "iluminan" cuando un sistema de IA está mintiendo. Esto puede ser mucho más manejable (incluso si da garantías menos sólidas).

En los últimos años, ha habido una oleada de trabajo emocionante en esta área. *CCS*es capaz de identificar una "dirección de la verdad" en modelos con solo datos no supervisados.⁹² *ROMA*es capaz de identificar la parte del modelo que sabe que la Torre Eiffel está en París y, a continuación, editar directamente el conocimiento del modelo para situar la Torre Eiffel en Roma. *La ingeniería de representación y las intervenciones en tiempo de inferencia* demuestran el uso de técnicas de arriba hacia abajo para detectar mentiras y alucinaciones y controlar quirúrgicamente el comportamiento del modelo en jailbreaking, búsqueda de poder, justicia, veracidad y más. Hay otros trabajos creativos sobre la detección de mentiras que ni siquiera requieren componentes internos del modelo.

⁹² Esencialmente, necesitando sólo el con- Es decir, que las etiquetas de verdad son las de verdadero/falso, que no tendremos para los sistemas sobrehumanos.

Soy cada vez más optimista en cuanto a que las técnicas de interpretabilidad de arriba hacia abajo serán una herramienta poderosa, es decir, podremos construir algunas algo así como un "detector de mentiras de IA"⁹³—y sin necesidad de funda-⁹³ Todavía me preocupa la escalabilidad avances mentales en la comprensión de las redes neuronales.

flagrantes.

Interpretabilidad de la cadena de pensamiento. Como se mencionó anteriormente, creo que es bastante plausible que iniciemos nuestro camino hacia AGI con sistemas que "piensen en voz alta" a través de cadenas de pensamiento. Aunque no es el algoritmo más eficiente en el límite (me sorprendería mucho si la superinteligencia todavía usara la cadena de pensamiento inglesa, en lugar de pensar a través de algún tipo de estados internos recurrentes), es posible que aún nos beneficiemos de esta propiedad para los primeros AGI. Esto sería una bendición increíble para la interpretabilidad: ¡tendríamos acceso al "monólogo interno" del AGI! Eso haría que fuera relativamente fácil detectar fallas de alineación

Sin embargo, hay mucho trabajo por hacer aquí, si quisiéramos confiar en esto. ¿Cómo nos aseguramos de que el CoT siga siendo legible? (Es

de muchas de estas técnicas a modelos muy per- humanos, creo que explícita o implícitamente se basan principalmente en etiquetas de verdad sobre el terreno, es decir, un supervisor más inteligente que el modelo, y/o generalización favorable.

puede simplemente pasar de un inglés comprensible a un lenguaje de modelos ininteligible, dependiendo de cómo usemos, por ejemplo, RL para entrenar modelos: ¿podemos agregar algunas restricciones simples para asegurarnos de que siga siendo legible?) ¿Cómo nos aseguramos de que el CdT sea *fiel*, es decir, que refleje exactamente lo que piensan los modelos? (Por ejemplo, hay **algunos trabajos** que muestran que, en ciertas situaciones, los modelos simplemente inventarán un razonamiento posthoc en su CoT que en realidad no refleja su razonamiento interno real para una respuesta).

Mi mejor conjetura es que una simple medición de legibilidad y fidelidad, y algunos trucos simples para preservar la legibilidad y la fidelidad por más tiempo, podrían llegar bastante lejos. Sí, esto no funcionará en algunos mundos, y es un truco un poco simple, pero es una fruta al alcance de la mano; En mi opinión, esta dirección está infravalorada desde el punto de vista penal.

TE CONFRONTATORIO TE Picadura Y MEDIDAS. A LO LARGO DEL camino, será fundamental realizar pruebas de estrés de la alineación de nuestros sistemas en cada paso: nuestro objetivo debe ser cumplir con cada falla en el laboratorio antes de que lo encontremos en la naturaleza. Esto requerirá un avance sustancial en las técnicas para el **trabajo en equipo rojo automatizado**. Por ejemplo, si colocamos deliberadamente puertas traseras o desalineaciones en los modelos, ¿nuestra capacitación en seguridad los habría detectado y eliminado de ellos? (Los primeros trabajos sugieren que "**agentes durmientes**" pueden sobrevivir a través de la formación en seguridad, por ejemplo).

En términos más generales, será fundamental tener buenas mediciones de alineación. ¿Tiene el modelo el poder de desalinearse? Por ejemplo, ¿tiene objetivos a largo plazo y qué tipo de impulsos está aprendiendo? ¿Y cuáles son las "líneas rojas" claras? Por ejemplo, un límite muy intuitivo podría ser "el modelo de razonamiento (cadena de pensamientos) siempre tiene que permanecer legible y fiel". (Como Eric Schmidt **dice**, el punto en el que los agentes de IA pueden hablar en un idioma que no podemos entender, debemos desconectar las computadoras). Otra podría ser desarrollar mejores mediciones para determinar si los modelos están siendo completamente honestos.

La ciencia de la medición de la alineación aún está en pañales;

Mejorar esto será fundamental para ayudarnos a hacer las concesiones correctas sobre el riesgo durante la explosión de inteligencia. Haciendo la ciencia

que nos permite medir la alineación y nos da una comprensión de "¿qué evidencia sería suficiente para asegurarnos de que la próxima OOM en territorio sobrehumano es segura?" es uno de los trabajos de mayor prioridad para la investigación de la alineación en la actualidad (más allá de solo un trabajo que trata de extender RLHF aún más a "algoritmos sobrehumanos").

Véase también [Este artículo sobre la superalineación](#) [Líneas de investigación](#) para la convocatoria de propuestas de Subvenciones Rápidas de Superalineación.

Automatización de la investigación de alineación

En última instancia, vamos a tener que automatizar la búsqueda de alineación. No hay forma de que logremos resolver directamente la alineación de la verdadera superinteligencia; Cubrir esa vasta brecha de inteligencia parece extremadamente difícil. Además, al final de la explosión de inteligencia, después de que 100 millones de investigadores de IA automatizada hayan superado furiosamente una década de progreso en el aprendizaje automático, espero que haya muchos más sistemas alienígenas en términos de arquitectura y algoritmos en comparación con el sistema actual (con propiedades potencialmente menos benignas, por ejemplo, sobre la legibilidad del CoT, las propiedades de generalización o la gravedad de la desalineación inducida por el entrenamiento).

Pero tampoco tenemos que resolver este problema por nuestra cuenta. Si logramos alinear sistemas algo sobrehumanos lo suficiente como para confiar en ellos, estaremos en una posición increíble: tendremos millones de investigadores de IA automatizados, más inteligentes que los mejores investigadores de IA, a nuestra disposición. Será decisivo aprovechar adecuadamente este ejército de investigadores autoacoplados para resolver la alineación de sistemas aún más sobrehumanos.

(Esto se aplica de manera más general, por cierto, para todas las especificaciones de los riesgos de la IA, incluido el uso indebido, etc. La mejor ruta, tal vez la única, hacia la seguridad de la IA en todos estos casos, implicará aprovechar adecuadamente los primeros AGI para la seguridad; por ejemplo

Deberíamos poner a un montón de ellos a trabajar en la investigación automatizada para mejorar la seguridad contra los actores extranjeros que exfiltran pesos, otros en el refuerzo de las defensas contra los peores ataques biológicos, y así sucesivamente).

Conseguir la alineación automatizada correcta durante la explosión de inteligencia será extraordinariamente arriesgado: pasaremos por muchos años de avances de IA en cuestión de meses, con poco

tiempo humano para tomar las decisiones correctas, y comenzaremos a entrar en un territorio donde las fallas de alineación podrían ser catastróficas. Los laboratorios deben estar dispuestos a dedicar una gran parte de su cómputo a la investigación automatizada de alineaciones (frente a la investigación automatizada de capacidades) durante la explosión de inteligencia, si es necesario.

Necesitaremos garantías sólidas que nos permitan confiar en la investigación automatizada de la alineación que se está produciendo y en mediciones de desalineación mucho mejores que las que tenemos hoy en día para saber si todavía estamos a salvo. Para cada OOM que queramos ascender, necesitaremos tener una confianza extremadamente alta en nuestros enfoques de alineación para ese próximo OOM.

Sobre todo, hacerlo bien requerirá un grado extremo de competencia, seriedad y voluntad de hacer concesiones difíciles. Una vez que nos acercamos a la verdadera superinteligencia, lo que está en juego se vuelve muy real, y el fracaso (la IA deshonesto) podría ser catastrófico.

Habrà momentos con evidencia ambigua, en los que los datos superficialmente se verán bien, pero hay algunas señales de advertencia que insinúan un peligro inminente, y en su mayoría ya no tenemos ni idea de lo que está pasando. Navegar esta situación de manera responsable requerirá hacer concesiones muy costosas. Eventualmente, la decisión de dar luz verde a la próxima generación de superinteligencia tendrá que tomarse tan en serio como la decisión de lanzar una operación militar.

Superdefensa

"Hacer bien la alineación" debería ser sólo el primero de muchos laicos de defensa durante la explosión de inteligencia. La alineación será difícil; *Habrà* fracasos en el camino. Si es posible, tenemos que estar en una posición en la que la alineación pueda fallar, pero el fracaso no sería catastrófico. Esto podría significar:

- *Seguridad*. Un cúmulo aislado es la primera capa de defensa contra la superinteligencia que intenta autoexfiltrarse y causar daño en el mundo real. Y eso es solo el comienzo; Necesitaremos una seguridad mucho más extrema contra el modelo de auto-

exfiltración⁹⁴ en todos los ámbitos, desde el cifrado de hardware hasta ⁹⁴Un modelo que roba sus propios pesos,

Aprobación de muchas claves.⁹⁵ Y una vez que nuestros AGI puedan hacer

para hacer copias de sí mismo fuera del centro de datos original.

⁹⁵protegiendo contra los humanos

engañados o persuadidos por la IA para ayudarla a exfiltrarse

Investigación y desarrollo, también deberíamos ponerlos a trabajar para reforzar aún más la seguridad.

- *Monitorización.* Si nuestros sistemas de IA están tramando algo sospechoso o malévolo, o si un empleado deshonesto intenta usarlos para actividades no autorizadas, debemos ser capaces de detectarlo. Sin embargo, esta no es una tarea trivial: tendremos millones de instituciones de AGI ejecutando nuestro clúster, ejecutando código complejo, etc. Necesitaremos sistemas avanzados de monitoreo que aprovechen los modelos en sí mismos, incluido el monitoreo de los resultados del modelo, las cadenas de pensamiento, el código escrito por el modelo, la actividad del servidor/red y las interacciones con investigadores / otros humanos; esencialmente, aprovechando algunos sistemas de IA para atacar otros sistemas de IA. Además, queremos desarrollar y emplear **sofisticados protocolos de "control"**, utilizando modelos menos poderosos pero confiables para protegernos contra la subversión de modelos más poderosos pero menos confiables. Esto puede requerir, por ejemplo, la misma cantidad de cómputo para supervisar las instancias del modelo que para las instancias del modelo activo.
- *Limitaciones de capacidad específicas.* En la medida de lo posible, debemos tratar de limitar las capacidades del modelo de manera específica que reduzcan las consecuencias de los errores. Un ejemplo central de esto podría ser eliminar todo lo relacionado con la biología y la química del entrenamiento de modelos (o usar técnicas de **"desaprendizaje"**); Esto ayuda a cortar los caminos hacia daños catastróficos en el mundo real que ocurren a través de armas biológicas y químicas
- *Restricciones de métodos de entrenamiento específicos.* Es probable que haya algunas formas de entrenar modelos que son inherentemente más riesgosas (más propensas a producir desalineaciones graves) que otras. Por ejemplo, el aprendizaje por imitación parece relativamente seguro (es difícil ver cómo eso conduciría a modelos que tienen objetivos internos peligrosos a largo plazo), mientras que deberíamos evitar el aprendizaje aleatorio basado en resultados a largo plazo (que parece mucho más probable que conduzca a al modelo de aprendizaje de metas indeseables a largo plazo) el mayor tiempo posible. También debemos evitar el entrenamiento contra nuestros métodos de interpretación y nuestra configuración de monitoreo (de modo que nuestros modelos no estén simplemente entrenados para trabajar alrededor de estos).

Puede haber limitaciones adicionales deseables, por ejemplo, mantener cadenas de pensamiento legibles y fieles siempre y cuando

posible. Debemos definir estas restricciones con anticipación, mantenerlas todo el tiempo que podamos durante la explosión de inteligencia y solo eliminarlas si es absolutamente necesario.

- Y es probable que haya mucho más posible aquí.

¿Serán infalibles? De nada. Es probable que la verdadera superinteligencia sea capaz de eludir la mayoría de los esquemas de seguridad, por ejemplo. Aun así, nos compran mucho más margen de error, y vamos a necesitar cualquier margen que podamos obtener. Querremos usar ese margen para llegar a una posición en la que tengamos una confianza muy alta en nuestras técnicas de alineación, solo relajando las medidas de "superdefensa" (por ejemplo, desplegando la superinteligencia en entornos no aislados) concomitantemente con nuestra confianza.

Las cosas volverán a complicarse una vez que pasemos a implementar estos sistemas de IA en entornos menos controlados, por ejemplo, en aplicaciones militares. Es probable que las circunstancias nos obliguen a hacer esto con bastante rapidez, pero siempre debemos tratar de comprar la mayor cantidad posible de información para el error, por ejemplo, en lugar de simplemente desplegar directamente las superinteligencias "en el campo" con fines militares, usándolas para hacer investigación y desarrollo de una manera más aislada.

y solo desplegar las tecnologías específicas que inventan (por ejemplo, sistemas de armas autónomas más limitados en los que estamos más seguros de que podemos confiar).

Por qué soy optimista y por qué tengo miedo

Soy increíblemente OPTIMISTA sobre la manejabilidad técnica del problema de la superalineación. Se siente como si hubiera toneladas de fruta al alcance de la mano en todas partes en el campo. En términos más generales, las realidades empíricas del aprendizaje profundo se han inclinado más a nuestro favor en comparación con lo que algunos especulaban hace 10 años. Por ejemplo, el aprendizaje profundo *se generaliza* de manera sorprendentemente benigna en muchas situaciones: a menudo simplemente "hace lo que queríamos decir" en lugar de recogiendo un comportamiento maligno abstruso.⁹⁶ Por otra parte, si bien ⁹⁶ Aunque, por supuesto, si bien eso puede

Comprender completamente los aspectos internos del modelo

será difícil, al menos para los

AGI iniciales, tenemos una oportunidad decente de interpretabilidad: podemos hacer que razonen de manera transparente a través de cadenas de pensamiento, y

Si se aplica a los modelos actuales y tal vez a los sistemas a nivel humano, tenemos que tener cuidado al intentar extrapolar la evidencia de los modelos actuales a futuros modelos enormemente sobrehumanos.

Las técnicas de piratería, como la ingeniería de representación, funcionan sorprendentemente bien, así como los "detectores de mentiras" o similares.

Creo que hay una posibilidad bastante razonable de que "el plan predeterminado"

Alinear sistemas "algo sobrehumanos" funcionará en su mayoría.⁹⁷ ⁹⁷ Para ser claros, dado lo que está en juego, creo que

Por supuesto, una cosa es hablar de un "plan predeterminado" en abstracto, y otra es que el equipo responsable de ejecutarlo El plan es tú y tus 20 compañeros (¡mucho más estresante!).⁹⁸

Todavía hay un número increíblemente pequeño de personas que trabajan seriamente en la solución de este problema, tal vez unas pocas docenas de investigadores serios. ¡Nadie está al tanto! Hay mucha investigación interesante y productiva de ML que hacer sobre esto, y la gravedad del desafío exige un esfuerzo mucho más concertado que el que estamos realizando actualmente.

Pero esa es solo la primera parte del plan: lo que realmente me quita el sueño es la explosión de inteligencia. Alinear los primeros AGIs, los primeros sistemas algo sobrehumanos, es una cosa. Vastamente sobrehumana, la superinteligencia alienígena es un nuevo juego de pelota, y es un *juego de pelota aterrador*.

La explosión de inteligencia se parecerá más a dirigir una guerra que a lanzar un producto. No estamos en camino a la superdefensa, a un cúmulo aislado ni nada de eso; No estoy seguro de que nos diéramos cuenta si un modelo se autofiltrara. No vamos por buen camino

Para que una cadena de mando sensata tome cualquiera de estas decisiones de alto riesgo, para insistir en la confianza muy alta apropiada para la superinteligencia, para tomar las decisiones difíciles de tomarse un tiempo extra antes de lanzar la próxima carrera de entrenamiento para lograr la seguridad correcta o dedicar una gran mayoría de la computación a la investigación de alineación, para reconocer el peligro que se avecina y evitarlo en lugar de chocar directamente contra él. En este momento, ningún laboratorio ha demostrado mucha voluntad de hacer concesiones costosas para lograr la seguridad correcta (tenemos muchos comités de seguridad, sí, pero esos son bastante insignificantes). Por defecto, es probable que nos topemos con la explosión de la inteligencia y hayamos pasado por unos cuantos OOM antes de que la gente se dé cuenta de en qué nos hemos metido.

Contamos demasiado con la suerte.

"Salir del paso" es, en cierto sentido, un plan terrible. Pero podría ser todo lo que tenemos.

⁹⁸ Me recuerda a Scott Aaronson

Carta a su yo más joven:

"Hay una empresa que está construyendo una IA que llena habitaciones gigantes, consume la electricidad de una ciudad y recientemente ha adquirido una asombrosa capacidad para conversar como la gente. Puede escribir ensayos o poesía sobre cualquier tema. Puede superar los exámenes de nivel universitario. Está adquiriendo diariamente nuevas capacidades de las que los ingenieros que se ocupan de la IA ni siquiera pueden hablar en público todavía. Esos ingenieros, sin embargo, se sientan

en la cafetería de la empresa y debatir el significado de lo que están creando.

¿Qué aprenderá a hacer la próxima semana? ¿Qué trabajos podría dejar obsoletos?

¿Deben reducir la velocidad o detenerse, por lo que

¿Para no hacerle cosquillas a la cola del dragón? Pero, ¿no significaría eso que alguien más, probablemente alguien con menos escrúpulos, despertaría primero al dragón?

¿Existe la obligación ética de contarle al mundo más sobre esto? ¿Hay obligación de contarle menos?

Yo —tú estás— pasando un año trabajando en esa empresa. Mi trabajo, tu trabajo, es desarrollar una teoría matemática sobre cómo evitar que la IA y sus sucesores causen estragos. Donde "causar estragos" podría significar cualquier cosa, desde impulsar la propaganda y el engaño académico, hasta dispensar consejos sobre bioterrorismo, hasta, sí, destruir el mundo".

IIId. El mundo libre debe prevalecer

La superinteligencia dará una ventaja económica y militar decisiva. China aún no está fuera de juego. En la carrera hacia el AGI, la supervivencia misma del mundo libre estará en juego. ¿Podemos mantener nuestra preeminencia sobre los poderes autoritarios? ¿Y lograremos evitar la autodestrucción en el camino?

La historia de la raza humana es la guerra, salvo breves y precarios interludios, nunca ha habido paz en el mundo; Y antes de que comenzara la historia, la lucha asesina era universal e interminable.

...

¿No podría encontrarse una bomba no más grande que una naranja que posea un poder secreto para destruir una manzana entera de edificios?

¿Concentrar la fuerza de mil toneladas de cordita y hacer estallar un corregimiento de un solo golpe?

WINSTON Churchill
"¿Nos suicidamos todos?" (1924)

La superinteligencia SERÁ la tecnología más poderosa y el arma más poderosa que la humanidad haya desarrollado jamás. Dará una ventaja militar decisiva, tal vez sólo comparable con las armas nucleares. Los autoritarios podían usar la superinteligencia para conquistar el mundo y para imponer el control total internamente. Los estados canallas podrían usarlo para amenazar con la aniquilación.

Y aunque muchos los descartan, una vez que el PCCh se despierte a la AGI, tendrá un camino claro para ser competitivo (al menos hasta que mejoremos drásticamente la seguridad de los laboratorios de IA de EE. UU.).

Cada mes de plomo también será importante para la seguridad. Nos enfrentamos a la

mayores riesgos si estamos enfrascados en una carrera reñida, con aliados demócratas y competidores autoritarios que corren a través de la ya precaria explosión de inteligencia a un ritmo vertiginoso, forzados

Dejar de lado cualquier precaución, temiendo que el otro obtenga primero la superinteligencia. Sólo si preservamos una ventaja saludable de aliados democráticos tendremos el margen de error para navegar por el período extraordinariamente volátil y peligroso en torno al surgimiento de la superinteligencia. Y sólo el liderazgo estadounidense es un camino realista para desarrollar un régimen de no proliferación que evite los riesgos de autodestrucción.

Nuestra generación da por sentado con demasiada facilidad que vivimos en paz y libertad. Y aquellos que anuncian la era de la AGI en la ciencia ficción con demasiada frecuencia ignoran el elefante en la habitación: la superinteligencia es un asunto de seguridad nacional, y Estados Unidos debe ganar.

Quienquiera que dirija la superinteligencia tendrá una ventaja militar decisiva

La superinteligencia no es cualquier otra tecnología (misiles hipersónicos, sigilo, etc.) en la que el liderazgo de Estados Unidos y de las democracias liberales es muy deseable, pero no estrictamente necesario. El equilibrio de poder militar puede mantenerse si Estados Unidos se queda atrás en una o un par de tecnologías de este tipo; Estas tecnologías son muy importantes, pero pueden ser superadas por las ventajas en otras áreas.

El advenimiento de la superinteligencia nos pondrá en una situación no vista desde el advenimiento de la era atómica: aquellos que la tengan ejercerán un dominio total sobre aquellos que no la tengan.

Anteriormente he hablado del vasto poder de la superinteligencia. Significará tener miles de millones de científicos, ingenieros y técnicos automatizados, cada uno mucho más inteligente que los científicos humanos más inteligentes, inventando furiosamente nuevas tecnologías, día y noche. La aceleración del desarrollo científico y tecnológico será extraordinaria. A medida que la superinteligencia se aplica a la investigación y el desarrollo en tecnología militar, podríamos pasar rápidamente por décadas de progreso tecnológico militar.

La Guerra del Golfo, o: Lo que unas pocas décadas de plomo tecnológico implican para el poder militar

La Guerra del Golfo proporciona una ilustración útil de cómo una ventaja de 20 a 30 años en tecnología militar puede ser decisiva. En ese momento, Irak comandaba el cuarto ejército más grande del mundo. En términos numéricos (tropas, tanques, artillería), la coalición liderada por Estados Unidos apenas fue igualada (o superada) por los iraquíes, mientras que los iraquíes habían tenido tiempo suficiente para atrincherar sus defensas (una situación que normalmente requeriría una ventaja de 3:1, o 5:1, en mano de obra militar para dislocarse).

Pero la coalición liderada por Estados Unidos *aniquiló* al ejército iraquí en una guerra terrestre de apenas 100 horas. Recuento de muertos de la coalición apenas 292, en comparación con los 20.000-50.000 iraquíes muertos y los cientos de miles de heridos o capturados. La Coalición perdió apenas 31 tanques, en comparación con la destrucción de más de 3.000 tanques iraquíes.

La diferencia en la tecnología no era divina ni infativa, pero era total y completamente decisiva: municiones guiadas e inteligentes, primeras versiones de sigilo, mejores sensores, mejores visores de tanques (para ver más lejos en la noche y en las tormentas de polvo), mejores aviones de combate, una ventaja en el reconocimiento, etc.

(Para un ejemplo más reciente, recordemos que Irán lanzó un ataque masivo de 300 misiles contra Israel, "el 99%" de los cuales fueron interceptados por la defensa antimisiles superior de Israel, Estados Unidos y sus aliados).

Una ventaja de uno o dos o tres años en la superinteligencia podría significar una ventaja militar tan decisiva como la que tuvo la coalición estadounidense contra Irak en la Guerra del Golfo. Estará en juego una remodelación completa del equilibrio de poder militar.

Imagínense si hubiéramos pasado por los desarrollos tecnológicos militares del siglo XX en menos de una década. Habríamos pasado de caballos, fusiles y trincheras, a modernos ejércitos de tanques, en un par de años; a armadas de aviones de combate supersónicos y armas nucleares y misiles balísticos intercontinentales un par de años después

ese; al sigilo y la precisión que pueden noquear a un enemigo antes de que sepan que estás allí un par de años después de eso.

Esa es la situación a la que nos enfrentaremos con el advenimiento de la superinteligencia: los avances tecnológicos militares de un siglo comprimidos a menos de una década. Veremos hackeos sobrehumanos que pueden paralizar gran parte de la fuerza militar de un adversario, roboejércitos y enjambres de drones autónomos, pero lo que es más importante, paradigmas completamente nuevos que aún no podemos empezar a imaginar, y la invención de nuevas armas de destrucción masiva con miles de veces más de poder destructivo (y también nuevas defensas de armas de destrucción masiva, como la defensa antimisiles impenetrable, que rápida y repetidamente trastornan los equilibrios de disuasión).

Y no sería solo el progreso tecnológico. A medida que resolvamos la robótica, la mano de obra se automatizará por completo, lo que también permitirá una explosión industrial y económica más amplia. Es plausible que las tasas de crecimiento puedan llegar a los 10 por ciento anuales; en una década, como máximo, el PIB de los que llevan la delantera superaría a los que van detrás. La rápida multiplicación de las fábricas de robots significaría no solo una drástica ventaja tecnológica, sino también una capacidad de producción para dominar el material puro. Pensemos en millones de receptores de misiles; miles de millones de drones; y así sucesivamente.

Por supuesto, no conocemos los límites de la ciencia y las muchas fricciones que podrían ralentizar las cosas. Pero no se necesitan avances divinos para obtener una ventaja militar decisiva. Y mil millones de científicos superinteligentes podrán hacer *mucho*. *Parece claro que en cuestión de años, los ejércitos anteriores a la superinteligencia se verían irremediabilmente superados.*

La ventaja militar sería decisiva incluso contra la disuasión nuclear

Para ser aún más claros: parece probable que la ventaja conferida por la superinteligencia sea lo suficientemente decisiva incluso como para eliminar preventivamente la disuasión nuclear de un adversario.

La mejora de las redes de sensores y el análisis podrían localizar incluso los submarinos nucleares actuales más silenciosos (de manera similar para los lanzadores de misiles móviles). Millones o miles de millones del tamaño de

un ratón

Los drones autónomos, con avances en sigilo, podrían infiltrarse detrás de las líneas enemigas y luego localizar, sabotear y decapitar subrepticamente las fuerzas nucleares del adversario. La mejora de los sensores, la selección de objetivos, etc., podría mejorar drásticamente la defensa antimisiles (similar a, por ejemplo, el ejemplo anterior de Irán contra Israel); Además, si se produce una explosión industrial, las fábricas de robots podrían producir miles de interceptores por cada misil contrario. Y todo esto sin siquiera considerar paradigmas científicos y tecnológicos completamente nuevos (por ejemplo, **desactivar** todas las armas nucleares).

Simplemente no sería un concurso. Y no sólo no hay competencia en el sentido nuclear de "podríamos destruirnos mutuamente", sino ninguna competencia en términos de ser capaz de destruir el poder militar de un rival sin sufrir bajas significativas. Un par de años de liderazgo en la superinteligencia significaría un dominio completo.

Si se produce una rápida explosión de inteligencia, es plausible que una ventaja de apenas *meses* pueda ser decisiva: meses podría significar la diferencia entre los sistemas de IA de nivel humano y los sistemas de IA sustancialmente sobrehumanos. Tal vez poseer esas superinteligencias iniciales por sí solas, incluso antes de ser ampliamente desplegadas, sería suficiente para una ventaja decisiva, p. ej. a través de habilidades de piratería sobrehumanas que podrían acabar con los ejércitos anteriores a la superinteligencia, enjambres de drones más limitados que amenazan con la muerte instantánea de cada líder opositor, funcionario y sus familias, y armas biológicas avanzadas desarrolladas con simulación al estilo AlphaFold que podrían apuntar a grupos étnicos específicos, por ejemplo, cualquier persona que no sea china Han (o simplemente retener la cura del adversario).

China puede ser competitiva

Muchos parecen complacientes con respecto a China y AGI. Los controles de exportación de chips los han neutralizado, y los principales laboratorios de IA están en los EE. UU. y el Reino Unido, por lo que no tenemos mucho de qué preocuparnos, ¿verdad? Los LLM chinos están *bien*, ¡definitivamente son capaces de entrenar modelos grandes!, pero son, en el mejor de los casos, comparables a los

segundo nivel de laboratorios de EE. UU.⁹⁹ E incluso los modelos chinos suelen ser ⁹⁹ Por ejemplo, Yi-Large parece ser

meras copias de los lanzamientos de código abierto estadounidenses (por ejemplo, la arquitectura Yi-34B parece tener esencialmente el Llama 2

arquitectura, con sólo un **Pocas líneas de código cambiadas**¹⁰⁰). Chi-Qwen abrazando

El aprendizaje profundo **solía ser** más importante de lo que es hoy en día (por ejemplo, **Baidu publicó** uno de los primeros artículos de derecho de escalado moderno), y aunque China **publica más artículos** sobre IA que Estados Unidos, no parece haber impulsado ninguno de los avances clave en los últimos años.

un modelo de clase GPT-4 en LMSys, pero eso es más de un año después de que OpenAI lanzara GPT-4.

¹⁰⁰ Del mismo modo, el código de la cara de

cita mucho a Mistral, y parece que la dependencia china de LLM del código abierto estadounidense es una **preocupación explícita** que ha llegado al primer ministro chino.

Sin embargo, todo eso no es más que un preludio. En caso de que el PCCh se despierte ante el AGI, deberíamos esperar esfuerzos extraordinarios por parte del PCCh para competir. Y creo que hay un camino bastante claro para que China esté en el juego: construir más que Estados Unidos y robar los algoritmos.

1. calcular

1a. Chips: China parece haber demostrado ahora la capacidad de **fabricar chips de 7 nm**. Si bien ir más allá de los 7 nm será difícil (lo que requiere EUV), ¡7 nm es suficiente! Como referencia, 7nm es lo que usó Nvidia A100s. El autóctono Huawei Ascend El 910B, basado en la plataforma SMIC de 7 nm, parece ser solo ~ 2-3 veces peor en rendimiento / \$ que un chip Nvidia equivalente ser.¹⁰¹

El **rendimiento** de la producción de 7 nm de SMIC y la madurez general de las capacidades chinas aquí se debaten,¹⁰² y una pregunta abierta crítica es en qué cantidades podrían producir estos chips de 7 nm.¹⁰³ Aun así, parece que hay al menos una posibilidad muy razonable de que puedan hacer esto a gran escala en unos pocos años.

¹⁰¹ El Huawei Ascend 910B parece **cuestan alrededor de 120,000 yuanes** por tarjeta, o alrededor de \$ 17k. Esto **se produce en el** mucha confianza en este momento, por ejemplo, de que no podrían gastar un poco más y obtener un amplio cómputo para los \$ 100B + y el billón de dólares

La mayor parte de las ganancias en chips de IA han provenido de la mejora del diseño de los chips adaptándolos a los casos de uso de IA (y es probable que China ya robe los diseños de chips de Nvidia de la cadena de suministro de Taiwán¹⁰⁴). 7 nm frente a 3 nm o 2 nm, y su inmadurez de fabricación general, probablemente encarece las cosas para China.¹⁰⁵ Pero eso no parece fatal de ninguna manera; se puede hacer una IA muy buena chips en la parte superior de un proceso de 7 nm. No tendría

SMIC de 7 nm, mientras se comporta de forma similar a un A100. Los H100 son quizás ~ 3 veces mejores que los A100, mientras que cuestan algo más (\$ 20-25k ASP), lo que sugiere solo un aumento de costo de ~ 2-3 veces para el rendimiento equivalente de la GPU AI para China en este momento.

¹⁰² Por ejemplo, todavía están usando la memoria HBM occidental (que por alguna razón no está controlada por exportación), aunque se dice que CXMT probará HBM el próximo año.

¹⁰³ Sin embargo, dado que todavía pueden importar otros tipos de chips de Occidente, podrían simplemente dirigir la totalidad de su nodo de 7 nm a chips de IA, compensando la menor producción general.

¹⁰⁴ Cabe destacar que incluso los ciberdelincuentes fueron capaces de hackear Nvidia y obtener secretos de diseño de GPU. Además, los diseños de TPUv 6 aparentemente se encontraban entre los robados por el ciudadano chino recientemente acusado en Google.

¹⁰⁵ Por ejemplo, tal vez estos sean 2 veces peores en perf/\$ o perf/Watt. A su vez, eso también significa que para lograr el mismo rendimiento general del centro de datos, necesita más potencia y necesita más chips conectados en red, lo que también hace que las cosas sean más complicadas.

formación en pocos años.¹⁰⁶

1b. La restricción vinculante para los clústeres de capacitación más grandes no serán los chips, sino la movilización industrial, quizás sobre todo los 100 GW de energía para el clúster de un billón de dólares. Pero si hay algo que China puede hacer mejor que Estados Unidos es *construir cosas*.

En la última década, China ha construido aproximadamente tanta nueva capacidad de electricidad como *toda la capacidad de Estados Unidos* (mientras que la capacidad de Estados Unidos se ha mantenido básicamente plana). En los EE. UU., estas cosas se atascan en la revisión ambiental, los permisos y la regulación durante una década. Por lo tanto, parece bastante plausible que China sea capaz de simplemente superar a los EE.UU. en los grupos de entrenamiento más grandes.

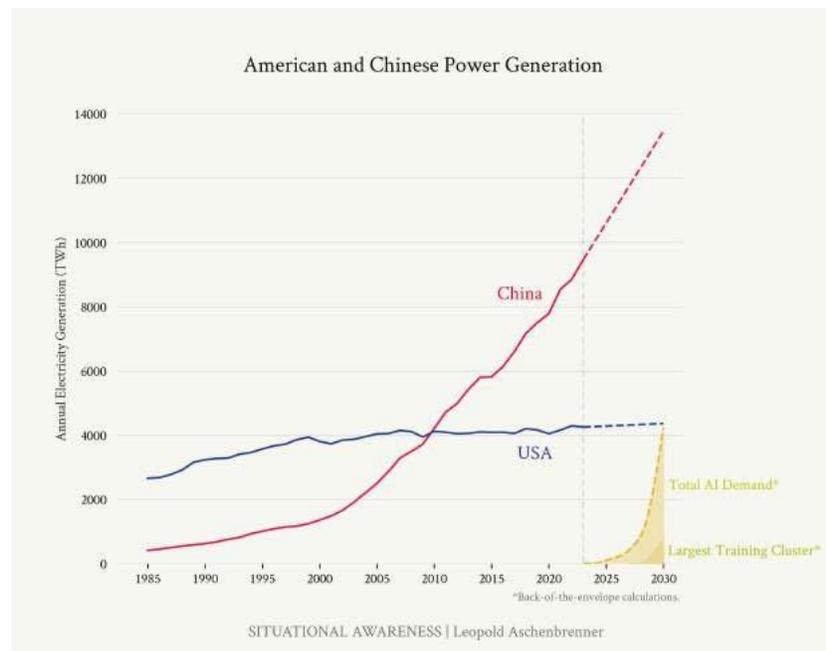


Figura 37: El aumento del poder de la IA para 2030 parece mucho más factible para China que para Estados Unidos. Basado en estimaciones anteriores de IIIa. Corriendo hacia el grupo del billón de dólares

2. ALGORITMOS

Como se discutió extensamente en Counting the OOMs, el escalado de la computación es solo una parte de la historia: los avances algorítmicos probablemente contribuyan al menos a la mitad del progreso de la IA. Estamos desarrollando los avances algorítmicos clave para AGI *en este momento* (esencialmente el EUV de los algoritmos debido a la pared de datos).

¹⁰⁶ Tenga en cuenta que incluso 3 veces más en fichas sería mucho menos que eso en términos de aumento en los costos de los centros de datos. Real

Los costos de fabricación lógica son <5% del costo de la GPU de Nvidia, e incluso considerando la memoria y CoWoS, es menos del 20% del precio de Nvidia debido a su margen. E incluso las propias GPU tienden a costar solo el 50-60% del costo de un centro de datos. Por lo tanto, 3 veces más caro en el extremo de la fabricación de chips podría traducirse en mucho, mucho menos que un aumento de 3 veces en el costo total. Incluso para chips 10 veces más caros, parece que China podría soportarlo sin aumentar enormemente los costos de los centros de datos.

Por defecto, espero que los laboratorios occidentales estén muy por delante; Tienen gran parte del talento clave, y en los últimos años han desarrollado todos los avances clave. El tamaño de la ventaja bien puede ser equivalente a un grupo 10 veces (o incluso 100 veces) más grande en unos pocos años; esto proporcionaría a los Estados Unidos una ventaja razonablemente cómoda.

Y, sin embargo, en el curso actual, renunciaremos por completo a esta ventaja: como se discutió ampliamente en la sección de seguridad, el estado actual de la seguridad esencialmente hace que sea trivial para China infiltrarse en los laboratorios estadounidenses. Por lo tanto, a menos que cerremos los laboratorios muy pronto, espero que China pueda simplemente robar los ingredientes algorítmicos clave para el AGI e igualar a EE. UU. Capacidades.¹⁰⁷
China robó

(Peor aún, si no mejoramos la seguridad, hay un camino aún más destacado para que China compita. Ni siquiera necesitarán entrenar su propio AGI: solo podrán robar las pesas AGI directamente. Una vez que hayan robado una copia del investigador de IA automatizado, se lanzarán a las carreras y podrán lanzar su propia explosión de inteligencia. Si están dispuestos a aplicar menos cautela —tanto una buena precaución como una regulación y demora irrazonables— que los EE.UU., podrían correr a través de la explosión de inteligencia más rápidamente, superándonos a nosotros en la superinteligencia).

Hasta la fecha, las empresas tecnológicas estadounidenses han hecho una apuesta mucho mayor por la IA y el escalado que cualquier esfuerzo chino; En consecuencia, estamos muy adelantados. Pero descartar a China ahora es un poco como descartar a Google en la carrera de la IA cuando salió ChatGPT a finales de 2022. Google aún no había centrado sus esfuerzos en una apuesta intensa por la IA, y parecía que OpenAI estaba muy por delante, pero una vez que Google despertó, un año y medio después, están dando una pelea muy seria. China también tiene un camino claro para poner en marcha un
Pelea muy seria. Si el PCCh se moviliza en la carrera hacia el AGI, el panorama podría empezar a ser muy diferente.

¹⁰⁷ Algunos argumentan que incluso si

estos secretos, no podrían competir porque requiere

Tal vez el gobierno chino sea incompetente; tal vez decidan que la IA amenaza al PCCh e impongan una regulación asfixiante. Pero no contaría con ello.

conocimiento. No estoy de acuerdo. Pienso en esto como si tuviera dos capas. La capa inferior es la destreza de ingeniería para carreras de entrenamiento a gran escala; Estas carreras de entrenamiento pueden ser delicadas y requieren conocimiento tácito. Pero como discutiré más adelante, los esfuerzos chinos de IA han demostrado ser perfectamente capaces de entrenar modelos a gran escala, y creo que tendrán este conocimiento tácito de manera autóctona. La capa superior es la receta algorítmica, cosas como la arquitectura del modelo, las leyes de escalado correctas, etc., que podrían transmitirse en una llamada de una hora. Estos multiplicadores de cómputo suelen ser cambios concretos, es decir, el conocimiento tácito subyacente para la formación a gran escala

Las ejecuciones deben transferirse. No creo que el "conocimiento tácito" sea una barrera decisiva para los esfuerzos chinos de AGI.

Yo, por mi parte, creo que tenemos que operar bajo el supuesto de que nos enfrentaremos a un esfuerzo chino de AGI a todo pulmón. Como cada año obtenemos saltos dramáticos en la capacidad de IA, a medida que comenzamos a ver la automatización temprana de los ingenieros de software, a medida que los ingresos de la IA se disparan y comenzamos a ver valoraciones de 10 billones de dólares y construcciones de clústeres de un billón de dólares, a medida que comienza a formarse un consenso más amplio de que estamos en la cúspide de AGI: el PCCh tomará nota. Por mucho que espere que estos saltos despierten al Gobierno de los Estados Unidos a la AGI, esperararía que despierten al PCCh a la AGI, y que se despierten a lo que significaría para su poder nacional estar atrasado en la AGI.

Serán un adversario formidable.

El peligro autoritario

Un dictador que ejerce el poder de la superinteligencia tendría un poder concentrado como nunca antes habíamos visto. Además de poder imponer su voluntad a otros países, A pesar de que la mayoría de las personas que se encuentran en el centro de la República, pueden consagrar su gobierno en el interior. Millones de agentes de la ley robóticos controlados por IA podrían vigilar a su población; la vigilancia masiva estaría hipercargada; Las IA leales a los dictadores podrían evaluar individualmente a cada ciudadano en busca de disidencia, con una detección avanzada de mentiras casi perfecta que erradicara cualquier disidencia.

Y lo que es más importante, la fuerza militar y policial robótica podría estar totalmente controlada por un solo líder político, y programada para ser perfectamente obediente, sin más riesgo de golpes de Estado o rebeliones populares.

Mientras que las dictaduras pasadas nunca fueron permanentes, la superinteligencia podía eliminar básicamente todas las amenazas históricas al gobierno de un dictador y asegurar su poder. Si el PCCh obtiene este poder, podría imponer la concepción de la "verdad" del Partido total y completamente.

Para ser claros, no solo me preocupa que los dictadores se vuelvan superinteligible porque "nuestros valores son mejores". Creo firmemente en la libertad y la democracia, porque no sé cuáles son los valores correctos. En el largo arco de la historia, "el tiempo ha trastornado a muchos

Religiones combatientes". Creo que debemos depositar nuestra fe en los mecanismos de corrección de errores, experimentación, competencia y adaptación.

La superinteligencia dará a quienes la ejerzan el poder de aplastar a la oposición, la disidencia y asegurar su gran plan de humanidad. Será difícil para cualquiera resistir la terrible tentación de usar este poder. Espero, sinceramente, que podamos confiar en la sabiduría de los Fundadores, permitiendo que florezcan valores radicalmente diferentes y preservando la estridente pluralidad que ha definido el experimento estadounidense.

Lo que está en juego en la carrera de AGI no será solo la ventaja en una guerra de poder de larga distancia, sino si la libertad y la democracia pueden sobrevivir durante el próximo siglo y más allá. El curso de la historia humana es tan brutal como claro. Dos veces en el siglo XX la tiranía amenazó al mundo; No debemos hacernos ilusiones de que esta amenaza está desterrada para siempre. Para muchos de mis jóvenes amigos, la libertad y la democracia se sienten como un hecho, pero no lo son. Con mucho, el sistema político más común de la historia es autoritarismo.¹⁰⁸

¹⁰⁸ Principalmente la monarquía.

Sinceramente, no conozco las intenciones del PCCh y sus aliados autoritarios. Pero, como recordatorio: el PCCh es un régimen fundado en la adoración continuada quizás el asesino en masa más talitario de la historia de la humanidad ("con estimaciones que oscilan entre 40 y 80 millones de víctimas debido al hambre, la persecución, el trabajo en las prisiones y las ejecuciones en masa"); un régimen que recientemente puso a un millón de uigures encampos de concentración y aplastó un Hong Kong libre; un régimen que practica sistemáticamente la vigilancia masiva para el control social, tanto del tipo novedoso (rastreo de teléfonos, bases de datos de ADN, reconocimiento facial, etc.) como del tipo anticuado (reclutar a un ejército de ciudadanos para informar sobre sus vecinos); un régimen que garantiza que todos los mensajes de texto pasen por un censor, y que va tan lejos como para reprimir la disidencia hasta el punto de llevar a las familias a las comisarías de policía cuando su hijo en el extranjero asiste a una protesta; un régimen que ha cimentado Xi Jinping como dictador vitalicio; un régimen que pregona sus objetivos de aplastar militarmente y "reeducar" a una nación vecina libre; un régimen que busca explícitamente un orden mundial centrado en China.

El mundo libre debe prevalecer sobre los poderes autoritarios en esta carrera. Debemos nuestra paz y libertad a la preeminencia económica y militar de Estados Unidos. Tal vez incluso dotado de superinteligencia, el PCCh se comporte de manera responsable en el escenario internacional, dejando a cada uno a su suerte. Pero la historia de los dictadores de su calaña no es bonita. Si Estados Unidos y sus aliados no logran ganar esta carrera, lo arriesgamos todo.

Mantener una ventaja saludable será decisivo para la seguridad

Es la historia maldita de la ciencia y la tecnología que, a medida que han desplegado sus maravillas, también han ampliado los medios de destrucción: desde palos y piedras, hasta espadas y lanzas, rifles y cañones, ametralladoras y tanques, bombarderos y misiles, armas nucleares. La curva "destrucción/\$" ha bajado constantemente a medida que la tecnología ha avanzado. Nosotros

Es de esperar que el rápido progreso tecnológico posterior a la superinteligencia siga esta tendencia.

Tal vez los avances dramáticos en biología produzcan nuevas armas biológicas extraordinarias, que se propaguen silenciosamente, rápidamente, antes de matar con perfecta letalidad cuando se les ordene (y que se pueden hacer extraordinariamente baratas, asequibles incluso para los grupos terroristas). Tal vez los nuevos tipos de armas nucleares permitan que el tamaño de los arsenales nucleares aumente en órdenes de magnitud, con nuevos mecanismos vectores que son indetectables. Tal vez drones del tamaño de un mosquito, cada uno con un veneno mortal, podrían ser el objetivo de matar a todos los miembros de una nación opuesta. Es difícil saber lo que produciría un siglo de progreso tecnológico, pero estoy seguro de que desplegaría posibilidades espantosas.

La humanidad apenas eludió la autodestrucción durante la Guerra Fría. Desde el punto de vista histórico, el mayor riesgo existencial planteado por la AGI es que nos permitirá desarrollar nuevos medios extraordinarios de muerte masiva. Esta vez, estos medios podrían incluso ser accesibles a actores deshonestos o terroristas (especialmente si, como en el curso actual, los pesos de la superinteligencia no están suficientemente protegidos y pueden ser robados directamente por Corea del Norte, Irán y compañía).

Corea del Norte ya tiene un programa concertado de armas biológicas: Estados Unidos **evalúa** que "Corea del Norte tiene un programa ofensivo dedicado a nivel nacional" para desarrollar y producir armas biológicas. Parece plausible que su principal limitación sea hasta qué punto su pequeño círculo de científicos de primer nivel ha sido capaz de empujar los límites de la biología (sintética). ¿Qué sucederá cuando se elimine esa restricción, cuando puedan usar millones de superinteligencias para acelerar su investigación y desarrollo de armas biológicas? Por ejemplo, Estados Unidos **considera** que Corea del Norte tiene actualmente una "capacidad limitada" para modificar genéticamente productos biológicos, ¿qué sucede cuando eso se vuelve ilimitado? ¿Con qué nuevos brebajes impíos nos tendrán como rehenes?

Además, como se discutió en la sección de superalineación, habrá riesgos extremos de seguridad alrededor y durante la explosión de inteligencia: nos enfrentaremos a nuevos desafíos técnicos para garantizar que podamos confiar y controlar de manera confiable los sistemas de IA sobrehumanos. Esto puede requerir que reduzcamos la velocidad en algunos momentos críticos, por ejemplo, retrasando 6 meses en medio de la explosión de inteligencia para obtener garantías adicionales sobre seguridad, o utilizando una gran fracción de la computación en la investigación de alineación en lugar del progreso de las capacidades.

Algunos esperan algún tipo de tratado internacional sobre seguridad. Esto me parece fantasioso. El mundo en el que tanto el PCCh como el Gobierno de los Estados Unidos están lo suficientemente controlados por la AGI como para tomarse en serio el riesgo de seguridad es también el mundo en el que ambos se dan cuenta de que está en juego el predominio económico y militar internacional, que estar meses atrasados en la AGI podría significar quedarse atrás permanentemente. Si la carrera está reñida, cualquier equilibrio en el control de armas, al menos en la fase inicial en torno a la superinteligencia, parece extremadamente inestable. En resumen, la "ruptura" es demasiado fácil: el incentivo (y el temor de que otros actúen sobre este incentivo) para avanzar con una explosión de inteligencia, para alcanzar la superinteligencia y la ventaja decisiva, demasiado grande.¹⁰⁹ Por lo menos, las probabilidades de que obtengamos algo lo suficientemente bueno aquí parecen escasas. (¿Cómo han ido esos tratados sobre el clima? Eso parece un problema dramáticamente más fácil en comparación con esto).

La principal —tal vez la única— esperanza que tenemos es que una alianza de democracias tenga una ventaja saludable sobre las potencias adversarias. El

¹⁰⁹ Considere la siguiente comparación de equilibrios de control de armas inestables vs. estables. El control de armamentos de la década de 1980 durante la Guerra Fría redujo sustancialmente las armas nucleares, pero se centró en un equilibrio estable. Estados Unidos y la Unión Soviética todavía tenían miles de armas nucleares. MAD estaba asegurado, incluso si uno de los actores intentaba un programa de choque para construir más armas nucleares; y una nación rebelde podría tratar de construir algunas armas nucleares propias, pero no amenazar fundamentalmente a los EE.UU. o a la Unión Soviética con una superación.

Sin embargo, cuando el desarme se produce a niveles muy bajos de armas o se produce en medio de un rápido cambio tecnológico, el equilibrio es inestable. Un actor deshonesto o un infractor de tratados puede iniciar fácilmente un programa de choque y amenazar con-

La cuenta supera a la de los otros jugadores. Cero armas nucleares no sería un equilibrio estable; Del mismo modo, [este documento](#) presenta interesantes estudios de casos históricos (como las limitaciones de armamentos posteriores a la Primera Guerra Mundial y el Tratado Naval de Washington) en los que el desarme en situaciones dinámicas similares se estabilizó, en lugar de estabilizarse.

Si los meros meses de ventaja sobre AGI dieran una ventaja absolutamente decisiva, esto hace que el desarme estable en la IA sea igualmente difícil. Un advenedizo deshonesto o un infractor de tratados podría obtener una gran ventaja iniciando en secreto un programa de choque; La tentación sería demasiado grande para que cualquier tipo de arreglo fuera estable.

Estados Unidos debe liderar, y usar ese liderazgo para hacer cumplir las normas de seguridad en el resto del mundo. Ese es el camino que tomamos con las armas nucleares, ofreciendo asistencia sobre los usos pacíficos de la tecnología nuclear a cambio de un régimen internacional de no proliferación (respaldado en última instancia por el poder militar estadounidense), y es el único camino que se ha demostrado que funciona.

Quizás lo más importante es que una pista saludable nos da espacio para mejorar: la capacidad de "cobrar" partes de la pista, si es necesario, para obtener la seguridad correcta, por ejemplo, dedicando trabajo adicional a la alineación durante la explosión de inteligencia.

Los desafíos de seguridad de la superinteligencia se volverían *extremadamente* difíciles de manejar si se está en una carrera armamentista reñida. Un adelanto de 2 años frente a uno de 2 meses podría marcar fácilmente toda la diferencia. Si solo tenemos una ventaja de 2 meses, no tenemos ningún margen de seguridad. Ante el temor a la explosión de inteligencia del PCCh, es casi seguro que correríamos, sin restricciones, a través de nuestra propia explosión de inteligencia, lanzándonos hacia sistemas de IA mucho más inteligentes que los humanos en meses, sin ninguna capacidad de reducir la velocidad para tomar decisiones clave correctas, con todos los riesgos de que la superinteligencia salga mal que eso implica. Nos enfrentaríamos a una situación extremadamente volátil, ya que nosotros y el PCCh desarrollamos rápidamente una nueva tecnología militar extraordinaria que desestabilizó repetidamente la disuasión. Si nuestros secretos y pesos no están encerrados, incluso podría significar que una serie de otros estados canallas también están cerca, cada uno de ellos utilizando la superinteligencia para proporcionar su propio nuevo arsenal de superarmas de destrucción masiva. Incluso si apenas logramos salir por delante, probablemente sería una victoria pírrica; La lucha existencial habría llevado al mundo al borde de la autodestrucción total.

La superinteligencia se ve muy diferente si los aliados democráticos

Tener una ventaja saludable, digamos 2 años.¹¹⁰ Eso nos da el tiempo necesario- ¹¹⁰ Obsérvese que, dado el ya rápido

Es necesario navegar por la serie de desafíos sin precedentes a los que nos enfrentaremos en torno y después de la superinteligencia, y para estabilizar la situación.

cuando ofrecemos un acuerdo a China y otros adversarios. Sabrán que no van a ganar, y por lo tanto sabrán que su único

Si queda claro que Estados Unidos ganará de manera decisiva, es

ritmo de progreso de la IA en la actualidad, y el ritmo aún más rápido que deberíamos

Esperen una explosión de inteligencia, y la explosión tecnológica más amplia posterior a la superinteligencia, "incluso" una explosión de 2-

El año de ventaja sería una *gran* diferencia en las capacidades en el mundo posterior a la superinteligencia.

opción es sentarse a la mesa; y preferimos evitar un enfrentamiento febril o intentos militares desesperados de su parte para sabotear los esfuerzos occidentales. A cambio de garantizar la no injerencia en sus asuntos y compartir los beneficios pacíficos de la superinteligencia, puede nacer un régimen de no proliferación, normas de seguridad y una apariencia de estabilidad post-superinteligencia.

En cualquier caso, a medida que profundizamos en esta lucha, no debemos olvidar la amenaza de la autodestrucción. Que lo logremos

la Guerra Fría en una sola pieza implicó demasiada suerte¹¹¹ —y la destrucción podría ser mil veces más potente que la que enfrentamos entonces. Un liderazgo saludable de una coalición de democracias liderada por Estados Unidos, y un ejercicio solemne de este liderazgo para estabilizar cualquier situación volátil en la que nos encontremos, es probablemente el camino más seguro para navegar más allá de este precipicio. Pero en el fragor de la carrera de AGI, es mejor que no la estropeemos.

¹¹¹ Daniel Ellsberg. Cuenta este río. Como uno de los planificadores de la guerra nuclear en RAND y en el aparato de seguridad nacional en ese momento.

La superinteligencia es un asunto de seguridad nacional

Está claro: la AGI es un desafío existencial para la seguridad nacional de los Estados Unidos. Es hora de empezar a tratarlo como tal.

Poco a poco, el Gobierno de los Estados Unidos está empezando a moverse. Los controles a la exportación de chips estadounidenses son un gran problema, y fueron una medida increíblemente profética en ese momento. Pero tenemos que ponernos serios en todos los ámbitos.

Estados Unidos *tiene* una ventaja. Solo tenemos que mantenerlo. Y lo estamos arruinando ahora mismo. Sobre todo, debemos cerrar rápida y radicalmente los laboratorios de IA, antes de que filtremos avances clave de AGI en los próximos 12 a 24 meses (o los propios pesos de AGI). Debemos construir los clústeres de cómputo en los EE.UU., no en dictaduras que ofrecen dinero fácil. Y sí, los laboratorios de IA estadounidenses tienen el deber de trabajar con la comunidad de inteligencia y el ejército. El liderazgo de Estados Unidos en AGI no garantizará la paz y la libertad simplemente creando las mejores aplicaciones de novia de IA. No es bonito, pero debemos construir IA para la defensa estadounidense.

Estamos en camino de la situación internacional más intensa de las últimas décadas. Putin está en marcha en Europa del Este. Oriente Medio está en llamas. El PCCh considera que tomar Taiwán como su destino. Ahora agregue la carrera a AGI. Añádase un siglo de avances tecnológicos comprimidos en años posteriores a la superinteligencia. Será una de las situaciones internacionales más inestables que se hayan visto jamás, y al menos inicialmente, la

Los incentivos para los primeros strikes serán tremendos.¹¹²

Ya hay una convergencia espeluznante de las líneas de tiempo de AGI (~ ¿2027?) y las líneas de tiempo de la invasión de Taiwán de los observadores de Taiwán (¿China está lista para invadir Taiwán para 2027?), una convergencia que seguramente solo se intensificará a medida que el mundo despierte a AGI. (¡Imagínese si en 1960, la gran mayoría de los depósitos de uranio del mundo estuvieran de alguna manera concentrados en Berlín!) Me parece que hay una posibilidad real de que el final del juego de AGI se desarrolle con el telón de fondo de la guerra mundial. Entonces todas las apuestas están canceladas.

¹¹² Cada uno correrá por su cuenta superinteligencias, y habrá una Ventana limitada antes de que alguien se adelante

se habrá alejado irreversiblemente. Habrá un gran incentivo para tratar de desactivar los grupos de superinteligencia enemigos antes de que hayan obtenido una ventaja física suficiente (por ejemplo, usar la superinteligencia para desarrollar una defensa antimisiles impenetrable o enjambres de drones) que deje a todos los demás permanentemente en el polvo.

IV. El proyecto

A medida que se intensifique la carrera hacia la AGI, el estado de seguridad nacional se involucrará. El Gobierno de los Estados Unidos despertará de su letargo, y para el 27/28 tendremos algún tipo de proyecto AGI del gobierno. Ninguna startup puede manejar la superinteligencia. En algún lugar de un SCIF, el final del juego estará en marcha.

Debemos tener curiosidad por saber cómo un conjunto de objetos de este tipo —cientos de centrales eléctricas, miles de bombas, decenas de miles de personas apiñadas en establecimientos nacionales— puede remontarse a unas pocas personas sentadas en bancos de laboratorio discutiendo el comportamiento peculiar de un tipo de átomo.

Spencer R. DESGASTADO

Hoy en día se presentan muchos planes para la "gobernanza de la IA", desde la concesión de licencias de sistemas de IA de frontera hasta las normas de seguridad y una nube pública con unos pocos cientos de millones de dólares en computación para los académicos. Estos parecen bien intencionados, pero a mí me parece que están cometiendo un error de categoría.

*Me parece una propuesta **descabellada** que el gobierno de EE.UU. permita que una startup de ciencia ficción al azar desarrolle superinteligencia. Imagínense si hubiéramos desarrollado bombas atómicas dejando que Uber improvisara.*

La superinteligencia —sistemas de IA *mucho* más inteligentes que los humanos— tendrá un gran poder, desde el desarrollo de armamento novedoso hasta la explosión del crecimiento económico. La superinteligencia será

el lugar de la competencia internacional; una ventaja de meses potencialmente decisiva en un conflicto militar.

Es una ilusión de aquellos que han interiorizado inconscientemente nuestro breve respiro de la historia que esto no convocará fuerzas más primordiales. Al igual que muchos científicos antes que nosotros, las grandes mentes de San Francisco esperan poder controlar el destino del demonio que están dando a luz. En este momento, todavía pueden; porque son de los pocos que tienen conciencia situacional, que comprenden lo que están construyendo. Pero en los próximos años, el mundo despertará. Lo mismo ocurrirá con el estado de seguridad nacional. La historia hará un regreso triunfal.

Como en muchas ocasiones anteriores —Covid, Segunda Guerra Mundial—, parecerá que Estados Unidos está dormido al volante, antes de que, de repente, el gobierno se ponga en marcha de la manera más extraordinaria.

Habrà un momento —dentro de unos años, solo un par de saltos más "a nivel de 2023" en las capacidades de los modelos y el discurso de la IA— en el que quedará claro: estamos en la cúspide de la AGI y de la superinteligencia poco después. Si bien hay mucho flujo en la mecánica exacta, de una forma u otra, el USG estará al timón; los laboratorios líderes se fusionarán ("voluntariamente"); Congreso se apropiará de billones para chips y energía; Se formó una coalición de democracias.

Las startups son excelentes para muchas cosas, pero una startup por sí sola simplemente no está equipada para estar a cargo del proyecto de defensa nacional más importante de los Estados Unidos. Necesitaremos la participación del gobierno para tener siquiera una esperanza de defendernos contra la amenaza total de espionaje a la que nos enfrentaremos; los esfuerzos privados de IA bien podrían estar entregando directamente superinteligencia al PCCh. Necesitaremos que el gobierno garantice aunque sea una apariencia de una cadena de mando sana; no se puede tener directores ejecutivos aleatorios (o juntas directivas de organizaciones sin fines de lucro al azar) con el botón nuclear. Necesitaremos que el gobierno gestione los graves desafíos de seguridad de la superinteligencia, que gestione la niebla de guerra de la explosión de inteligencia. Necesitaremos que el gobierno despliegue la inteligencia para defenderse de cualquier amenaza extrema que se desarrolle, para superar la situación internacional extraordinariamente volátil y desestabilizada que seguirá. Necesitaremos que el gobierno

A pesar de que la mayoría de las personas que se encuentran en el centro de la Unión Europea se han convertido en una coalición para ganar la contienda con la mayoría de las personas que se han convertido

autoritarias, y forjar (y aplicar) un régimen de no proliferación para el resto del mundo. Ojalá no fuera así, pero necesitaremos al gobierno. (Sí, independientemente de la Administración).

En cualquier caso, mi principal afirmación no es normativa, sino descriptiva. En unos años, The Project estará en marcha.

El camino hacia el proyecto

Un giro de los acontecimientos grabado a fuego en mi memoria es de finales de febrero a mediados de marzo de 2020. En esas últimas semanas de febrero y primeros días de marzo, estaba en una desesperación total: parecía claro que estábamos en el covid-exponencial: una plaga estaba a punto de barrer el país, el colapso de nuestros hospitales era inminente y, sin embargo, casi nadie lo tomó en serio. El alcalde de Nueva York seguía desestimando los temores de Covid como racismo y animando a la gente a ir a los espectáculos de Broadway. Lo único que podía hacer era comprar mascarillas y vender en corto el mercado.

Y, sin embargo, en unas pocas semanas, todo el país se cerró y el Congreso se había apropiado de billones de dólares (literalmente >10% del PIB). Ver hacia dónde podría ir el exponencial antes de tiempo era demasiado difícil, pero cuando la amenaza se acercó lo suficiente, lo suficientemente existencial, se desataron fuerzas extraordinarias. La respuesta fue tardía, cruda, contundente, pero llegó, y fue dramática.

Los próximos años en IA se sentirán similares. Ahora estamos en la mitad del juego. 2023 ya fue un cambio salvaje. AGI pasó de ser un tema marginal con el que dudarías en asociarte, al tema de las audiencias del Senado y las cumbres de líderes mundiales. Teniendo en cuenta lo temprano que aún estamos, el nivel de compromiso del Gobierno de los Estados Unidos ha sido impresionante para mí. Un par de "2023" más, y la ventana de Overton se abrirá por completo.

A medida que avanzamos a través de los OOM, los saltos continuarán. Para 2025/2026 más o menos, espero los próximos cambios verdaderamente impactantes; La IA generará ingresos anuales de \$ 100B + para las grandes empresas de tecnología y superará a los doctores en inteligencia bruta para resolver

problemas. Al igual que

el colapso del mercado de valores de Covid hizo que muchos tomaran Covid en serio, tendremos empresas de \$ 10 billones y la manía de la IA estará en todas partes. Si eso no es suficiente, para 2027/28, tendremos modelos entrenados en el clúster de \$ 100 mil millones +; Los agentes de IA de pleno derecho/trabajadores remotos comenzarán a automatizar ampliamente la ingeniería de software y otros trabajos cognitivos. Cada año, la aceleración se sentirá vertiginosa.

Si bien muchos aún no ven la posibilidad de AGI, eventualmente se formará un consenso. Algunos, como Szilard, vieron la posibilidad de una bomba atómica mucho antes que otros. Su alarma no fue bien recibida inicialmente; La posibilidad de una bomba fue descartada como remota (o al menos, se consideró que lo más conservador y apropiado era minimizar la posibilidad). Las fervientes súplicas secretas de Szilard fueron objeto de burla e ignoración. Pero muchos científicos, inicialmente escépticos, comenzaron a darse cuenta de que una bomba era posible a medida que llegaban más y más resultados empíricos. Una vez que la mayoría de los científicos llegaron a creer que estábamos en la cúspide de una bomba, el gobierno, a su vez, consideró que la exigencia de seguridad nacional era demasiado grande, y el Proyecto Manhattan se puso en marcha.

A medida que las OOM pasen de la extrapolación teórica a la realidad empírica (extraordinaria), poco a poco, se formará también un consenso entre los principales científicos, ejecutivos y funcionarios gubernamentales: estamos en la cúspide, en la cúspide de la AGI, en la cúspide de una explosión de inteligencia, en la cúspide de la superinteligencia. Y en algún momento de aquí, tendremos las primeras demostraciones genuinamente aterradoras de la IA: tal vez la tan discutida "ayudar a los novatos a fabricar armas biológicas", o piratear sistemas críticos de forma autónoma, o algo completamente diferente. Quedará claro: esto La tecnología será una tecnología militar absolutamente decisiva. Incluso si tenemos la suerte de no estar en una guerra importante, parece probable que el PCCh se haya dado cuenta y haya lanzado un formidable esfuerzo de AGI. Tal vez el eventual (inevitable) descubrimiento de la infiltración del PCCh en los principales laboratorios de IA de Estados Unidos cause un gran revuelo.

Alrededor del 26/27 más o menos, el estado de ánimo en Washington se volverá sombrío. La gente empezará a sentir visceralmente lo que está ocurriendo; Estarán *asustados*.

Desde los pasillos del Pentágono hasta la trastienda, las sesiones informativas del Congreso sonarán lo obvio

pregunta, la pregunta en la mente de todos: ¿necesitamos un Proyecto AGI Manhattan? Poco a poco al principio, luego de repente, se irá aclarando: esto está pasando, las cosas se van a poner *salvajes*, este es el desafío más importante para la seguridad nacional de los Estados Unidos desde la invención de la bomba atómica. De una forma u otra, el estado de seguridad nacional se involucrará mucho. El proyecto será la respuesta necesaria, de hecho la única plausible.

Por supuesto, este es un relato extremadamente abreviado: mucho depende de cuándo y cómo se forma el consenso, los disparos de advertencia clave, etc. DC es infamemente disfuncional. Al igual que con el Covid, e incluso el Proyecto Manhattan, el gobierno llegará increíblemente tarde y torpe. Después de la carta de Einstein al presidente en 1939 (redactado por Szilard), se formó un Comité Consultivo sobre el Uranio. Pero los funcionarios eran incompetentes, y al principio no pasó mucho. Por ejemplo, Fermi solo recibió 6 mil dólares (alrededor de 135 mil dólares de hoy) para apoyar su investigación, e incluso eso no se le dio fácilmente y solo se recibió después de meses de espera. Szilard creía que el proyecto se había retrasado durante al menos un año por la miopía y lentitud de las autoridades. En marzo de 1941, el gobierno británico finalmente concluyó que una bomba era inevitable. Inicialmente, el comité de EE.UU. ignoró por completo este informe británico durante meses, hasta que finalmente, en diciembre de 1941, se lanzó un esfuerzo de bomba atómica a gran escala.

Hay muchas maneras de poner esto en práctica en la práctica. Para ser claros, esto no tiene por qué parecer una nacionalización literal, ya que los investigadores de los laboratorios de IA ahora son empleados por los militares o los investigadores de los laboratorios de IA. lo que sea (¡aunque podría!).¹¹³ Más bien, espero un orquestación. La relación con el Departamento de Defensa podría parecerse a la relación que el Departamento de Defensa tiene con Boeing o Lockheed Martin. Tal vez a través de la contratación de defensa o similar, se establece una empresa conjunta entre los principales proveedores de computación en la nube, los laboratorios de IA y el gobierno, lo que lo convierte funcionalmente en un proyecto del estado de seguridad nacional. Al igual que los laboratorios de IA se comprometieron "voluntariamente" con la Casa Blanca en 2023, los laboratorios occidentales podrían acordar más o menos "voluntariamente" fusionarse en el esfuerzo nacional. Y es probable que el Congreso

¹¹³ Tenga en cuenta que, si bien las

tenga que involucrarse, dada la

ayudar a desarrollar componentes para armas nucleares, nunca se les permite poseer un arma nuclear completa y ensamblada. En comparación, la versión principal del "proyecto de gobierno AGI" que estoy proponiendo aquí está **billones de inversión involucrados, y para controles y equilibrios.**¹¹⁴ Cómo se revelan todos estos detalles es una historia para otro día.

privatizada sin precedentes, para la clase de referencia de armas de destrucción masiva.

¹¹⁴ Congreso, ¡incluso el vicepresidente!— no sabía nada del Proyecto Manhattan. Probablemente no deberíamos repetir eso aquí; Incluso sugeriría que los funcionarios clave de The Project requieran la confirmación del Senado.

Pero a finales del 26/27/28 estará en marcha. El equipo central de investigación de AGI (unos pocos cientos de investigadores) se trasladará a un lugar seguro; El clúster de un billón de dólares se construirá a una velocidad récord; El proyecto estará en marcha.

¿Por qué el proyecto es el único camino?

No me hago ilusiones sobre el gobierno. Los gobiernos se enfrentan a todo tipo de limitaciones y escasos incentivos. Soy un gran creyente en el sector privado estadounidense, y casi nunca abogaré por una fuerte participación del gobierno en la tecnología o la industria.

Solía aplicar este mismo marco a AGI, hasta que me uní a un laboratorio de IA. Los laboratorios de IA son muy buenos en algunas cosas: han sido capaces de llevar la IA de un proyecto científico académico al gran escenario comercial, de una manera que solo una startup puede hacer. Pero, en última instancia, los laboratorios de IA siguen siendo startups. Simplemente no deberíamos esperar que las startups estén equipadas para manejar la superinteligencia.

No hay buenas opciones aquí, pero no veo otra manera. Cuando una tecnología se vuelve tan importante para la seguridad nacional, necesitaremos al Gobierno de los Estados Unidos.

La superinteligencia será el proyecto de defensa nacional más importante de los Estados Unidos

He discutido el poder de la superinteligencia en artículos anteriores. En cuestión de años, la superinteligencia sacudirá **por completo el equilibrio de poder militar**. A principios de la década de 2030, la integridad del arsenal estadounidense (nos guste o no, la base de la paz y la seguridad mundiales) probablemente será obsoleta. No será solo una cuestión de modernización, sino un reemplazo total.

En pocas palabras, quedará claro que el desarrollo de AGI caerá en una categoría más parecida a las armas nucleares que a Internet. Sí, por supuesto que será de doble uso, pero la tecnología nuclear también fue de doble uso. Las aplicaciones civiles tendrán su tiempo. Pero en

la niebla del final del juego de AGI, para bien o para mal, la seguridad nacional será el telón de fondo principal.

Tendremos que remodelar por completo las fuerzas estadounidenses, en cuestión de años, frente a los rápidos cambios tecnológicos, o arriesgarnos a ser completamente superados por los adversarios que sí lo hagan. Quizás, por encima de todo, la prioridad inicial será desplegar la superinteligencia para aplicaciones defensivas, para desarrollar contramedidas para sobrevivir a incontables nuevas amenazas: adversarios con capacidades de piratería sobrehumanas, nuevas clases de enjambres de drones sigilosos que podrían ejecutar un ataque preventivo contra nuestra disuasión nuclear. la proliferación de avances en biología sintética que pueden convertirse en armas, las turbulentas luchas de poder internacionales (y nacionales) y los proyectos de superinteligencia deshonestos.

Ya sea nominalmente privado o no, el proyecto AGI tendrá que ser, *será*, integralmente un proyecto de defensa, y requerirá una cooperación extremadamente estrecha con el estado de seguridad nacional.

Una cadena de mando sensata para la superinteligencia

El poder, y los desafíos, de la superinteligencia caerán en una clase de referencia muy diferente a cualquier otra cosa que estemos acostumbrados a ver en las empresas tecnológicas. Parece bastante claro: esto no debería estar bajo el mando unilateral de un CEO cualquiera. De hecho, en el mundo de los laboratorios privados que desarrollan superinteligencia, es bastante plausible que los directores ejecutivos individuales tengan el poder *para, literalmente, golpear al gobierno de Estados Unidos.*¹¹⁵ Imagínate si Elon Musk¹¹⁵ Ni siquiera requeriría tenía el mando final del arsenal nuclear.¹¹⁶ (O si una junta sin fines de lucro al azar pudiera decidir tomar el control del arsenal nuclear).

Tal vez sea obvio, pero: como sociedad, hemos decidido

los gobiernos cráticos deben controlar a los militares;¹¹⁷ superin- La telelengua será, al menos al principio, el arma militar más poderosa. La propuesta radical no es El Proyecto; la propuesta radical es apostar a que los directores ejecutivos privados de IA ejerzan el poder militar y se conviertan en dictadores benévolos.

(De hecho, en el mundo de los laboratorios privados de IA,

¹¹⁵ Imagínate si Elon Musk¹¹⁵ Ni siquiera requeriría A estas alturas, la mayoría de los empleados del laboratorio de IA se han automatizado en su mayor parte.

¹¹⁶ Y como dijo una vez Sam Altman, cada año que nos acerquemos a AGI todo el mundo ganará +10 puntos locos.

¹¹⁷ De hecho, el gobierno que ha probablemente sería incluso peor que los directores ejecutivos aleatorios con el botón nuclear, parte de la IA

¡Las armas más grandes fueron un enorme logro civilizatorio! En lugar de luchas medievales de todos contra todos, resolveríamos los desacuerdos a través de tribunales, instituciones pluralistas, etc.

La pésima seguridad de los laboratorios es su absoluta falta de controles internos. Es decir, los empleados aleatorios del laboratorio de IA (sin investigación) podrían pasar desapercibidos).

Necesitaremos una cadena de mando sensata, junto con todos los demás procesos y salvaguardas que necesariamente vienen con el manejo responsable de lo que será comparable a una ADM, y requerirá que el gobierno lo haga. En cierto sentido, se trata simplemente de un argumento burkeano: las instituciones, las constituciones, las leyes, los tribunales, los controles y equilibrios, las normas y la dedicación común al orden democrático liberal (por ejemplo, los generales que se niegan a seguir órdenes ilegales), etc., controlan el poder del gobierno han resistido la prueba de cientos de años. Especial Mientras tanto, las estructuras de gobernanza de los laboratorios de IA colapsaron la primera vez que se probaron. El ejército estadounidense ya podría matar a casi todos los civiles de Estados Unidos, o tomar el poder, si quisiera, y la forma en que mantenemos bajo control el poder del gobierno sobre las armas nucleares no es a través de muchas empresas privadas con sus propios arsenales nucleares. Sólo hay una cadena de empresas y un conjunto de instituciones que ha demostrado estar a la altura de esta tarea.

Una vez más, tal vez seas un verdadero libertario y no estés de acuerdo normativamente (deja que Elon Musk y Sam Altman comanden los suyos arsenales nucleares!).¹¹⁸ Pero una vez que se hace evidente que la superintel-

código abierto
La importancia es un asunto principal de seguridad nacional, estoy seguro de que así es como lo verán los hombres y mujeres de Washington.

¹¹⁸ O tal vez digas, solo

desarrollo de defensa más caro durante la Segunda Guerra Mundial, más caro que el Manhat-

Los usos civiles de la superinteligencia

Por supuesto, eso no significa que las aplicaciones civiles de la superinteligencia estén reservadas para el gobierno.

- La reacción nuclear en cadena se aprovechó primero como un proyecto del gobierno —y las armas nucleares se reservaron permanentemente para el gobierno—, pero la energía nuclear civil floreció como proyectos privados (en los años 60 y 70, antes de que los ecologistas la cerraran).
- Boeing fabricó el B-29 (el proyecto de investigación y

todo. El problema con el simple código abierto de todo es que no es un mundo feliz de mil flores floreciendo en los EE. UU., sino un mundo en el que el PCCh tiene libre acceso a la superinteligencia desarrollada en los EE. UU. y puede construir más que (y aplicar menos precaución/regulación) y apoderarse del mundo. Y el otro problema, por supuesto, es la proliferación de superarmas de destrucción masiva en todos los estados canallas y grupos terroristas del mundo. No creo que acabe bien. Es un poco como si no tener ningún gobierno condujera más a la tiranía (o a la destrucción) que a la libertad.

En cualquier caso, la gente sobrevalora la importancia del código abierto a medida que nos acercamos a AGI. Dados los costos de los clústeres que ascienden a cientos de miles de millones, y los secretos algorítmicos clave ahora son propietarios en lugar de publicarse como hace un par de años, serán 2 o 3 o más jugadores líderes, en lugar de una comunidad feliz de codificadores descentralizados que construyan AGI.

Creo que una variante diferente del código abierto seguirá desempeñando un papel importante: Modelos que se retrasan un par de años con respecto al código abierto, lo que ayuda a que los beneficios de la tecnología se difundan ampliamente.

y los bombarderos de largo alcance B-47 y B-52 en asociación con el ejército, antes de usar esa tecnología para el Boeing 707, el avión comercial que marcó el comienzo de la era de los reactores. Y hoy en día, aunque Boeing solo puede vender aviones de combate furtivos al gobierno, puede desarrollar y vender libremente aviones civiles de forma privada.

- Y así ocurrió con los radares, los satélites, los cohetes, la tecnología genética, las fábricas de la Segunda Guerra Mundial, etc.

El desarrollo inicial de la superinteligencia estará dominado por la exigencia de seguridad nacional de sobrevivir y estabilizar un período increíblemente volátil. Y los usos militares de la superinteligencia permanecerán reservados para el gobierno, y se harán cumplir las normas de seguridad. Pero una vez que el peligro inicial ha pasado, el camino natural es que las empresas involucradas en el consorcio nacional (y otras) se dediquen privadamente a las solicitudes civiles.

Incluso en mundos con El Proyecto, un ecosistema privado, pluralista, basado en el mercado y floreciente de aplicaciones civiles de la superinteligencia tendrá su día.

Seguridad

He hablado largo y tendido sobre esto en un artículo anterior en las series. En el curso actual, es mejor que renunciemos a cualquier esfuerzo de AGI estadounidense; China puede robar rápidamente todos los avances algorítmicos y los pesos de los modelos (literalmente, una copia de la superinteligencia) directamente. Ni siquiera está claro que lleguemos a una seguridad "a prueba de Corea del Norte" para la superinteligencia en el curso actual. En el mundo de las startups privadas que desarrollan AGI, la superinteligencia proliferaría en docenas de estados canallas. Es simplemente insostenible.

Si vamos a tomarnos esto en serio, obviamente tenemos que cerrar estas cosas. La mayoría de las empresas privadas no se han tomado esto en serio. Pero en cualquier caso, si finalmente vamos a enfrentarnos a toda la fuerza del espionaje chino (por ejemplo, robar las pesas es la prioridad # 1 del MSS), probablemente sea imposible que una empresa privada obtenga una seguridad lo suficientemente buena. Requerirá

amplia cooperación de la comunidad de inteligencia de EE.UU. en ese momento para asegurar suficientemente a AGI. Esto implicará restricciones invasivas en los laboratorios de IA y en el equipo central de investigadores de AGI, desde la investigación extrema hasta el monitoreo constante, el trabajo desde un SCIF hasta la reducción de la libertad para irse; y requerirá una infraestructura que solo el gobierno puede proporcionar, incluida en última instancia la seguridad física de los propios centros de datos de AGI.

En cierto sentido, la seguridad por sí sola es suficiente para necesitar el proyecto del gobierno: tanto la preeminencia del mundo libre como la seguridad de la IA están condenadas si no podemos bloquear estas cosas. (De hecho, creo que es bastante probable que sea un factor importante en el resultado final: una vez que la infiltración china en los laboratorios de AGI se haga evidente, todos los senadores y congresistas y funcionarios de seguridad nacional Voluntad de la Corte. . . tener una opinión firme sobre el asunto).

Seguridad

En pocas palabras: hay muchas maneras de estropear esto, desde asegurarnos de que podemos controlar y confiar de manera confiable en los miles de millones de agentes superinteligentes que pronto estarán a cargo de nuestra economía y ejército (el problema de la superalineación), hasta controlar los riesgos del mal uso [delos nuevos medios de destrucción masiva](#).

Algunos laboratorios de IA afirman estar comprometidos con la seguridad: reconocen que lo que están construyendo, si sale mal, podría causar catástrofe y prometen que harán lo que sea necesario cuando llegue el momento. No sé si podemos confiar en su promesa lo suficiente como para arriesgar la vida de todos los estadounidenses en ella. Lo que es más importante, hasta ahora no han demostrado la competencia, la confiabilidad o la seriedad necesarias para lo que ellos mismos reconocen que están construyendo.

En el fondo, son startups, con todos los incentivos comerciales habituales. La competencia podría empujarlos a todos a simplemente correr a través de la explosión de inteligencia, y al menos habrá algunos actores que estarán dispuestos a dejar la seguridad en el camino. En particular, es posible que queramos "gastar parte de nuestro liderazgo" para tener tiempo para resolver los desafíos de seguridad, pero los laboratorios occidentales deberán coordinarse para hacerlo. (Y, por supuesto,

a los laboratorios privados ya les habrán robado sus pesas AGI,
por lo que su seguridad pre-

Las precauciones ni siquiera importarán; estaremos a merced de las precauciones de seguridad del PCCh y de Corea del Norte).

Una respuesta es la regulación. Eso puede ser apropiado en mundos en los que la IA se desarrolla más lentamente, pero me temo que la regulación simplemente no estará a la altura de la naturaleza del desafío de la explosión de la inteligencia. Lo que se necesita no será tanto como pasar unos años haciendo evaluaciones cuidadosas e impulsando algunos estándares de seguridad a través de una burocracia. *Será más como pelear una guerra.*

Nos enfrentaremos a un año demencial en el que la situación está cambiando extremadamente rápido cada semana, en el que las decisiones difíciles basadas en datos ambiguos serán de vida o muerte, en el que las soluciones, incluso los problemas mismos, no estarán ni cerca de estar completamente claros de tiempo, sino que se reduce a la competencia en una "niebla de guerra", lo que implicará concesiones locas como "algunas de nuestras mediciones de alineación parecen ambiguas, ya no entendemos realmente lo que está pasando, podría estar bien, pero hay algunas señales de advertencia de que la próxima generación de superinteligencia podría salir mal, ¿deberíamos retrasar la próxima ejecución de entrenamiento por 3 meses para obtener más confianza en la seguridad, pero oh no, los últimos informes de inteligencia indican que China nos robó los pesos y está avanzando en su propia explosión de inteligencia, ¿qué debemos hacer?".

No estoy seguro de que un proyecto gubernamental sea competente para lidiar con esto, pero la alternativa de la "superinteligencia desarrollada por startups" parece mucho más cercana a "rezar por lo mejor" de lo que comúnmente se reconoce. Necesitaremos una cadena de comunicación que pueda poner sobre la mesa la seriedad que requerirá hacer estas difíciles concesiones.

Estabilización de la situación internacional

La explosión de inteligencia y sus secuelas inmediatas darán lugar a una de las situaciones más volátiles y tensas que la humanidad haya enfrentado jamás. Nuestra generación no está acostumbrada a esto. Pero en este período inicial, la tarea que tenemos entre manos no será la de construir productos frescos. Será de alguna manera, desesperadamente, superar este período.

Necesitaremos que el proyecto del gobierno gane la carrera contra el

poderes autoritarios— y para darnos el liderazgo claro y el respiro necesario para sortear los peligros de esta situación. Más vale que nos demos por vencidos si no podemos evitar el robo instantáneo de pesos de modelos de superinteligencia.

Querremos agrupar los esfuerzos de Occidente: reunir a nuestros mejores científicos, utilizar todas las GPU que podamos encontrar y garantizar que los billones de dólares de la construcción de clústeres se produzcan en los Estados Unidos. Tendremos que proteger los centros de datos contra el sabotaje de los adversarios o los ataques directos.

Quizás, sobre todo, se necesitará el liderazgo estadounidense para desarrollar —y si es necesario, hacer cumplir— un régimen de no proliferación. Haremos lo siguiente:

necesitan subvertir a Rusia, Corea del Norte, Irán y los grupos terroristas para que no usen su propia superinteligencia para desarrollar tecnología y armamento que les permita mantener al mundo como rehén.

Tendremos que usar la superinteligencia para reforzar la seguridad de nuestra infraestructura crítica, el ejército y el gobierno para defendernos de las nuevas capacidades extremas de piratería. Necesitaremos usar la superinteligencia para estabilizar el equilibrio entre ataque y defensa de los avances en biología o similares. Tendremos que desarrollar herramientas para controlar de forma segura la superinteligencia y para acabar con las inteligencias rebeldes que surgen de los proyectos poco cuidadosos de otros. Los sistemas de IA y los robots se moverán a una velocidad humana de 10-100x+; Todo empezará a suceder muy rápido. Tenemos que estar preparados para manejar cualquier otro trastorno de seis sigma. y las amenazas concomitantes— surgen de comprimir el progreso tecnológico de un siglo en unos pocos años.

Al menos en este período inicial, nos enfrentaremos a la más extraordinaria exigencia de seguridad nacional. Quizás, nadie esté preparado para esta tarea. Pero de las opciones que tenemos, The Project es la única sensata.

El proyecto es inevitable; si es bueno no lo es

En última instancia, MI AFIRMACIÓN PRINCIPAL aquí es descriptiva: nos guste o no, la superinteligencia no se verá como una startup de ciencia ficción y, de alguna manera, estará principalmente en el dominio de la seguridad nacional. He hablado mucho de The Project con mis amigos de San

Francisco en el último año. Quizás lo que más me ha sorprendido es

Qué sorprendida está la mayoría de la gente con la idea. Simplemente no han considerado la posibilidad. Pero una vez que lo consideran, la mayoría está de acuerdo en que parece obvio. Si estamos en lo cierto acerca de lo que creemos que estamos construyendo, *por supuesto*, al final esto será (de alguna forma) un proyecto de gobierno. Si un laboratorio desarrollara una superinteligencia literal mañana, *por supuesto* que los federales intervendrían.

Una variable libre importante no es *si*, sino *cuándo*. ¿Acaso el gobierno no se da cuenta de lo que está sucediendo hasta que estamos en medio de una explosión de inteligencia, o se dará cuenta un par de años antes? Si el proyecto de gobierno es inevitable, antes parece mejor. Necesitaremos mucho esos dos años para llevar a cabo el programa de seguridad contra fallas, para poner a los funcionarios clave al día y preparados, para construir un laboratorio fusionado que funcione, etc. Será mucho más caótico si el gobierno solo interviene al final (y los secretos y pesos ya habrán sido robados).

Otra variable gratuita importante es la coalición internacional que podemos reunir: tanto una alianza más estrecha de democracias para el desarrollo de la superinteligencia, como una oferta más amplia de participación en los beneficios hecha al resto del mundo.

- El primero podría parecerse al Acuerdo de **Quebec**: un pacto de secreto entre Churchill y Roosevelt para poner en común sus recursos para desarrollar armas nucleares, sin utilizarlas entre sí o contra otros sin un compromiso mutuo. Querremos incorporar al Reino Unido (Deepmind), a aliados de Asia Oriental como Japón y Corea del Sur (cadena de suministro de chips) y a la OTAN y otros aliados democráticos centrales (base industrial más amplia). Un esfuerzo conjunto tendrá más recursos, talento y controlar toda la cadena de suministro; permitir una estrecha coordinación en materia de seguridad, seguridad nacional y desafíos militares; y proporcionar controles y equilibrios útiles para ejercer el poder de la superinteligencia.
- Este último podría ser como **Átomos para la Paz**, el OIEA y el **TNP**. Deberíamos ofrecernos a compartir los beneficios pacíficos de la superinteligencia con un grupo más amplio de países (incluidos los no democráticos) y comprometernos a

no utilizar ofensivamente la superinteligencia contra ellos. A cambio, se abstienen de llevar a cabo sus propios proyectos de superinteligencia, hacen

compromisos de seguridad en el despliegue de sistemas de IA y aceptar restricciones a las aplicaciones de doble uso. La esperanza es que esta oferta reduzca los incentivos para la carrera armamentista y la proliferación, y reúna una amplia coalición bajo el paraguas de Estados Unidos para el orden mundial posterior a la superinteligencia.

Tal vez la variable libre más importante es simplemente si el inevitable proyecto de gobierno será competente. ¿Cómo se organizará? ¿Cómo podemos hacer esto? ¿Cómo funcionarán los controles y equilibrios, y cómo es una cadena de mando sana? Apenas se ha prestado atención a la idea de esto fuera.¹¹⁹ Casi todos los demás laboratorios de IA y la política de gobernanza de la IA son mis hermanos de Estudios del Progreso:

¹¹⁹ A

Un espectáculo secundario. Este es el juego de pelota.

acompañarnos en el viaje, será. . . estresante. Pero será nuestro deber servir a los libres

El final del juego

Y así, PARA la 27/28, el final del juego estará en marcha. Para el 28/29 la explosión de inteligencia estará en marcha; Para el año 2030, habremos convocado a la Superinteligencia, en todo su poder y poderío.

Quienquiera que pongan a cargo del Proyecto va a tener una tarea infernal: construir AGI, y construirlo rápido; para poner el La economía estadounidense en pie de guerra para fabricar cientos de millones de GPU; para bloquearlo todo, eliminar a los espías y defenderse de los ataques totales del PCCh; gestionar de alguna manera un millón de AGI automatizando furiosamente la investigación de la IA, haciendo saltos de una década en un año, y pronto produciendo sistemas de IA *mucho* más inteligentes que los humanos más inteligentes; para mantener las cosas lo suficientemente juntas como para que esto no se descarrile y produzca una superinteligencia deshonesto que intente arrebatar el control a sus supervisores humanos; utilizar esas superinteligencias para desarrollar cualquier nueva tecnología que sea necesaria para estabilizar la situación y mantenerse por delante de los adversarios, rehacer rápidamente las fuerzas estadounidenses para integrarlas; Todo ello mientras navega por lo que probablemente será la situación internacional más tensa jamás vista. Más vale que sean buenos, lo diré.

Para aquellos de nosotros que recibimos la llamada para

Si piensas en esto, ¡esta será la culminación de tu proyecto intelectual! Usted pasa todo este tiempo estudiando las instituciones de investigación del gobierno estadounidense, su decadencia en el último medio siglo y lo que se necesitaría para que volvieran a ser efectivas. Cuéntame: ¿cómo vamos a hacer efectivo el proyecto?

Figura 38: Oppenheimer y Groves.



y a toda la humanidad. Si lo logramos y miramos hacia atrás en esos años, será lo más importante que hayamos hecho. Y aunque cualquier instalación segura que encuentren probablemente no tendrá las comodidades del estilo de vida de los investigadores de IA ridículamente sobrecargado de hoy en día, no será tan malo. SF al- ready se siente como una peculiar ciudad universitaria de investigadores de IA; Probablemente esto no será tan diferente. Será el mismo círculo extrañamente pequeño que sudará las curvas de escala durante el día y pasará el rato durante el fin de semana, hablando sobre AGI y la política de laboratorio del día.

Excepto, bueno, lo que está en juego será

demasiado real. Nos vemos en el desierto,

amigos.

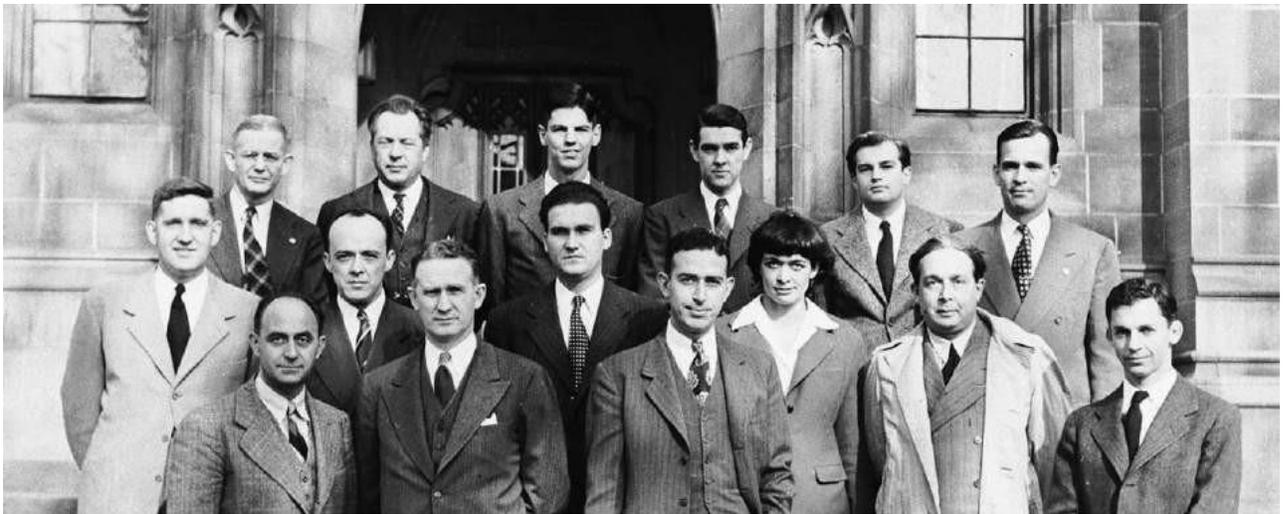


Figura 39: Reunión de científicos atómicos en el cuarto aniversario de la primera reacción de fisión nuclear controlada en la Universidad de California.

V. Pensamientos de despedida

¿Y si estamos en lo cierto?

Recuerdo la primavera de 1941 hasta el día de hoy. Entonces me di cuenta de que una bomba nuclear no solo era posible, sino que era inevitable.

Tarde o temprano estas ideas no podían ser peculiares para nosotros. Todo el mundo pensaría en ellos en poco tiempo, y algún país los pondría en acción.

Y no había nadie con quien hablar de eso, pasé muchas noches sin dormir. Pero sí me di cuenta de lo muy grave que podía ser. Y entonces tuve que empezar a tomar pastillas para dormir. Era el único remedio, nunca he parado desde entonces. Han pasado 28 años, y creo que no me he perdido ni una sola noche en todos esos 28 años.

James Chadwich

(Premio Nobel de Física y autor del informe del gobierno británico de 1941 sobre la inevitabilidad de una bomba atómica, que finalmente impulsó el Proyecto Manhattan a la acción)

Antes DE QUE TERMINE EL proceso , habremos construido la superinteligencia. De eso se ha tratado la mayor parte de esta serie. Para la mayoría de las personas con las que hablo en ciencia ficción, ahí es donde la pantalla se vuelve negra. Pero la década siguiente, la década de 2030, será al menos igual de agitada. Al final, el mundo se habrá transformado total e irreconociblemente. Se habrá forjado un nuevo orden mundial. Pero, ¡ay!, esa es una historia para otro momento.

Tenemos que llegar a un final, por ahora. Permítaseme hacer algunas observaciones finales.

Realismo AGI

Todo esto es mucho para contemplar, y muchos no pueden. "¡El aprendizaje profundo está chocando contra un muro!", proclaman cada año. Es solo otro boom tecnológico, dicen los expertos con confianza. Pero incluso entre los que están en el epicentro de la ciencia ficción, el discurso se ha polarizado entre dos gritos de guerra fundamentalmente poco serios.

Por un lado están los agoreros. Han estado observando la AGI durante muchos años; Les doy mucho crédito por su presciencia. Pero su pensamiento se ha osificado, desligado de las realidades empíricas del aprendizaje profundo, sus propuestas son ingenuas e inviables, y no logran comprometerse con ellas.

la amenaza autoritaria muy real. Las afirmaciones rabiosas de un 99% de probabilidades de fatalidad, los llamados a pausar indefinidamente la IA, claramente no son el camino.

En el otro extremo están los e/accs. En términos estrictos, tienen algunos puntos buenos: el progreso de la IA debe continuar. Pero debajo de su superficial publicación de en Twitter, son una farsa; diletantes que solo quieren construir sus nuevas empresas envoltorias en lugar de mirar a AGI a la cara. Dicen ser ardientes defensores de la libertad estadounidense, pero no pueden resistir el canto de sirena del desagradable dinero de los dictadores. En verdad, son verdaderos estancacionistas. En su intento para negar los riesgos, niegan el AGI; Esencialmente, todo lo que obtendremos son chatbots geniales, que seguramente no son peligrosos. (Ese es un aceleracionismo decepcionante en mi libro).

Pero tal y como yo lo veo, las personas más inteligentes del espacio han convergido en una perspectiva diferente, una tercera vía, a la que llamaré **Realismo AGI**. Los principios básicos son simples:

1. ***La superinteligencia es una MATERIA de seguridad nacional.*** Estamos construyendo rápidamente máquinas más inteligentes que los humanos más inteligentes. Este no es otro boom genial de Silicon Valley; No se trata de una comunidad aleatoria de programadores que escriben una Esto no es diversión ni juegos. La superinteligencia va a ser *salvaje*, será el arma más poderosa que la humanidad haya construido jamás. Y para cualquiera de nosotros involucrado, será lo más importante que hagamos en nuestra vida.

2. ***América MOSTO Conducir.*** La antorcha de la libertad no sobrevivirá a que Xi obtenga primero la AGI. (Y, siendo realistas, el liderazgo estadounidense también es el único camino hacia una AGI segura). Eso significa que no podemos simplemente "hacer una pausa"; significa que necesitamos ampliar rápidamente la producción de energía de los Estados Unidos para construir los clústeres AGI en los Estados Unidos. Pero también significa Amateur Startup Security entregando los secretos nucleares a la El PCCh ya no será suficiente, y significa que la infraestructura central de AGI debe ser controlada por Estados Unidos, no por algún dictador en Oriente Medio. Los laboratorios estadounidenses de IA deben anteponer el interés nacional.
3. ***No tenemos que meter la pata.*** Reconocer el poder de la superinteligencia significa también reconocer su peligro. Existen riesgos de seguridad muy reales; Riesgos muy reales Todo esto sale mal, ya sea porque la humanidad usa el poder destructivo que surge para nuestra aniquilación mutua, o porque, sí, la especie alienígena que estamos convocando es una que aún no podemos controlar por completo. Estos son manejables, pero improvisar no será suficiente. Para sortear estos peligros será necesario contar con buenas personas que aporten un nivel de seriedad que aún no se ha ofrecido.

A medida que la aceleración se intensifica, solo espero que el discurso se vuelva más estridente. Pero mi mayor esperanza es que haya quienes sientan el peso de lo que viene y lo tomen como un solemne llamado al deber.

¿Y si estamos en lo cierto?

Llegados a este punto, puede que pienses que yo y toda la gente de ciencia ficción estamos totalmente locos. Pero considere, solo por un momento: *¿y si tienen razón?* Estas son las personas que inventaron y construyeron esta tecnología; creen que la AGI se desarrollará en esta década; y, aunque hay un espectro bastante amplio, muchos de ellos se toman muy en serio la posibilidad de que el camino hacia la superinteligencia se desarrolle como he descrito en esta serie.

Es casi seguro que me he equivocado en partes importantes de la historia; Si la realidad resulta ser tan loca, las barras de error serán muy grandes. Además, como dije al principio, creo que hay un amplio abanico de posibilidades. Pero creo que es impor-

Es decir, que la mayoría de las personas que se encuentran en el centro de la ciudad de Y en esta serie he expuesto lo que actualmente creo que es el escenario más probable para el resto de la década, el resto de *esta* década.

Porque, está empezando a sentirse real, *muy* real. Hace unos años, al menos para mí, me tomé en serio estas ideas, pero eran abstractas, estaban en cuarentena en modelos y estimaciones de probabilidad. Ahora se siente extremadamente visceral. Puedo *verlo*. Puedo *ver* cómo se construirá AGI. Ya no se trata de estimaciones del tamaño del cerebro humano y extrapolaciones hipotéticas y teóricas y todo. Básicamente, puedo decirles el clúster en el que se entrenará AGI y cuándo se construirá, la combinación aproximada de algoritmos que usaremos, los problemas no resueltos y el camino para resolverlos, la lista de personas que importarán. Puedo *verlo*. Es extremadamente visceral. Claro, ir a por todas con Nvidia a principios de 2023 ha sido genial y todo, pero las cargas de la historia son pesadas. Yo no elegiría esto.

Pero lo más aterrador es que *no hay un equipo de crack que venga a manejar esto*. Cuando eres niño, tienes esta visión glorificada del mundo, que cuando las cosas se vuelven reales, están los científicos heroicos, los militares súper competentes, los líderes tranquilos que están en el mundo.

¿Quién salvará el día? No es así. El mundo es increíblemente pequeño; Cuando la fachada se cae, por lo general son solo unas pocas personas detrás de escena que son los jugadores en vivo, que están tratando desesperadamente de evitar que las cosas se desmoronen.

En este momento, hay quizás unos pocos cientos de personas en el mundo que se dan cuenta de lo que está a punto de golpearnos, que entienden lo locas que están a punto de ponerse las cosas, que tienen conciencia situacional. Probablemente conozco personalmente o estoy en un grado de separación de todos los que podrían dirigir plausiblemente el Proyecto. Las pocas personas detrás de escena que están tratando desesperadamente de evitar que las cosas se desmoronen son usted y sus amigos y sus amigos. Eso es todo. Eso es todo lo que hay.

Algún día estará fuera de nuestras manos. Pero en este momento, al menos durante los próximos años de la mitad del juego, el destino del mundo depende de estas personas.

¿Prevalecerá el mundo libre?

¿Domesticaremos a la superinteligencia o ella nos domesticará a nosotros? ¿Podrá la humanidad eludir la autodestrucción una vez más?

Lo que está en juego no es menor.

Son personas grandes y honorables. Pero son solo personas. Pronto, las IA estarán dirigiendo el mundo, pero nos espera un último rodeo. Que su mayordomía final traiga honor a la humanidad.

Apéndice

Algunos detalles adicionales sobre los cálculos de cómputo en la parte posterior de la envoltura

Proceso de entrenamiento

Año. El informe técnico de OpenAI GPT-4 afirmaba que GPT-4 finalizó el entrenamiento en agosto de 2022. A partir de entonces, avanzamos la tendencia aproximada de 0,5 OOM/año.

Equivalente a H100s. Semianálisis, JP Morgan y otros estiman que GPT-4 se entrenó en 25k A100, y los H100 son 2-3 veces el rendimiento de los A100.

Costar. A menudo, la gente cita cifras como "100 millones de dólares para el entrenamiento de GPT-4", utilizando sólo el coste de alquiler de las GPU (es decir, algo así como "¿cuánto costaría alquilar un clúster de este tamaño por 3 meses de formación?"). Pero eso es un error. Lo que importa es algo más parecido a ~el costo real de construir el clúster. Si quieres uno de los clústeres más grandes del mundo, ¡no puedes alquilarlo por 3 meses! Y, además, necesita el proceso para algo más que la ejecución de entrenamiento insignia: habrá muchos experimentos de reducción de riesgo, ejecuciones fallidas, otros modelos, etc.

Para aproximar el costo del clúster GPT-4:

- Las estimaciones públicas sugieren que el GPT-4 4 racimos que están alrededor 25k A100s.

- Suponiendo \$ 1 / A100 hora durante 2-3 años da aproximadamente un costo de \$ 500 millones.
- Alternativamente, puede estimarlo como un costo de \$ 25k por H100, equivalente a 10k H100 y GPU Nvidia que cuestan aproximadamente la mitad del costo de un clúster (el resto es energía, el centro de datos físicos, refrigeración, redes, personal de mantenimiento, etc.). (Por ejemplo, este [análisis del costo total de propiedad](#) estima que alrededor del 40% del costo de un clúster grande son las GPU H100 en sí, y otro 13% se destina a Nvidia para las redes de Infiniband. Dicho esto, excluir el costo de capital en ese cálculo significaría que las GPU son aproximadamente el 50% del costo, y con la conexión en red, Nvidia obtiene un poco más del 60% del costo del clúster).

FLOP/\$ está mejorando un poco para cada generación de Nvidia, pero no una tonelada. Por ejemplo, el H100 -> B100 es probablemente algo así como 1,5x FLOP/\$: los B100 son efectivamente dos H100 grapados juntos, pero se venden al por menor por <2 veces el costo. Pero el B100 era algo de una excepción. Son sorprendentemente baratos, probablemente porque Nvidia [quiere aplastar a la competencia](#). Por el contrario, los A 100 -> H100 no fueron una gran mejora de FLOP/\$ (2 veces mejor chip sin fp8, aproximadamente 2 veces el costo), tal vez 1.5 veces si contamos las mejoras de fp8, y eso fue para una generación de dos años.

Si bien creo que hay algunos vientos de cola para promover las mejoras de FLOP/\$ a partir de la compresión de márgenes, las GPU también podrían volverse más caras a medida que se limitan masivamente. Las ganancias de la especialización de los chips de IA continuarán, pero no me queda claro que todavía habrá mejoras técnicas que cambien las reglas del juego para FLOP/\$, dado que los chips ya están bastante especializados para la IA (por ejemplo, especializados para Transformers, y ya con precisión fp8/fp4), la Ley de Moore es glacial en estos días, y [otros componentes de cuello de botella como la memoria y la interconexión](#) están mejorando más lentamente. Si nos fijamos en [los datos de Epoch](#), parece que hay menos de 10 veces en FLOP/\$ en la última década para las principales GPU de ML, y por las razones antes mencionadas, en todo caso, esperararía que esto se ralentice.

Algo así como una mejora del 35%/año en FLOP/\$ nos daría el costo de 1T para el clúster de +4 OOM. Tal vez el FLOP/\$ mejore más rápido, pero también el gasto de capital del

centro de datos va a obtener más

Es caro, simplemente porque tendrá que construir nueva energía, haciendo una gran cantidad de gastos de capital por adelantado, en lugar de simplemente alquilar plantas de energía depreciadas existentes.

De todos modos, estos son números muy aproximados. Estaría muy dentro de las barras de error si, por ejemplo, el clúster 1T se puede hacer de manera más eficiente y en realidad produce más como +4.5 OOM en cómputo.

Costar. Un H100 tiene 700 W, pero necesita un montón de energía para el centro de datos (refrigeración, redes, almacenamiento); Estimaciones de semianálisis ~ 1.400W por H100.

Hay algunas ganancias que se pueden obtener en FLOP/Watt, aunque una vez que hemos agotado las ganancias de la especialización de los chips de IA (véase la nota anterior), por ejemplo, bajando a la precisión más baja posible, estas parecen algo limitadas (en su mayoría solo mejoras en el proceso de los chips, que son lentas). Dicho esto, a medida que la potencia se convierte en una restricción (y, por lo tanto, en una fracción mayor de los costos), los diseños de chips podrían especializarse en una mayor eficiencia energética a costa de los FLOP. Aun así, sigue existiendo la demanda de energía para la refrigeración, la red, el almacenamiento, etc. (que en los números H100 anteriores ya era aproximadamente la mitad de la demanda de energía).

Para estos números del reverso del sobre, trabajemos con 1kW por equivalente a H100 aquí; Una vez más, estos son solo cálculos aproximados. (Y si se produjera un avance inesperado en FLOP/Watt, esperaríamos el mismo gasto de energía, sólo que mayores ganancias de cómputo OOM).

Clases de referencia de potencia. Un clúster de 10 GW en funcionamiento continuo durante un año es de 87,6 TWh. En comparación, Oregón consume alrededor de 27 TWh de electricidad al año, el estado de Washington consume alrededor de 92 TWh al año.

Un clúster de 100 GW en funcionamiento continuo durante un año es de 876 TWh, mientras que la producción total anual de electricidad de EE. UU. es de aproximadamente 4.250 TWh.

[Haga clic aquí](#) para volver a la tabla "Proceso de entrenamiento".

Proceso general

Equivalentes a H100 este año. Estimo que Nvidia enviará del orden de 5 millones de GPU para centros de datos en 2024. Una minoría de ellos son B100, que contaremos como 2x+ H100. Luego están los otros chips de IA: TPU, Trainium, silicio personalizado de Meta, GPU AMD, etc.

Capacidad de TSMC. TSMC tiene capacidad para más de 150.000 obleas de 5 nm al mes, está aumentando a 100.000 obleas de 3 nm al mes, y probablemente otras 150.000 obleas de 7 nm al mes; llamémoslo 400.000 obleas al mes en total.

Digamos aproximadamente 35 H100 por oblea (H 100 se fabrican en 5 nm). Con 5-10 millones de equivalentes a H100 en 2024, eso es de 150 mil a 300 mil obleas por año para la producción anual de chips de IA en 2024.

Dependiendo de dónde en ese rango y si queremos contar la producción de 7 nm, eso es alrededor del 3-10% de la producción anual de obleas de vanguardia.

[Haga clic aquí](#) para volver a la tabla de "cómputo general".