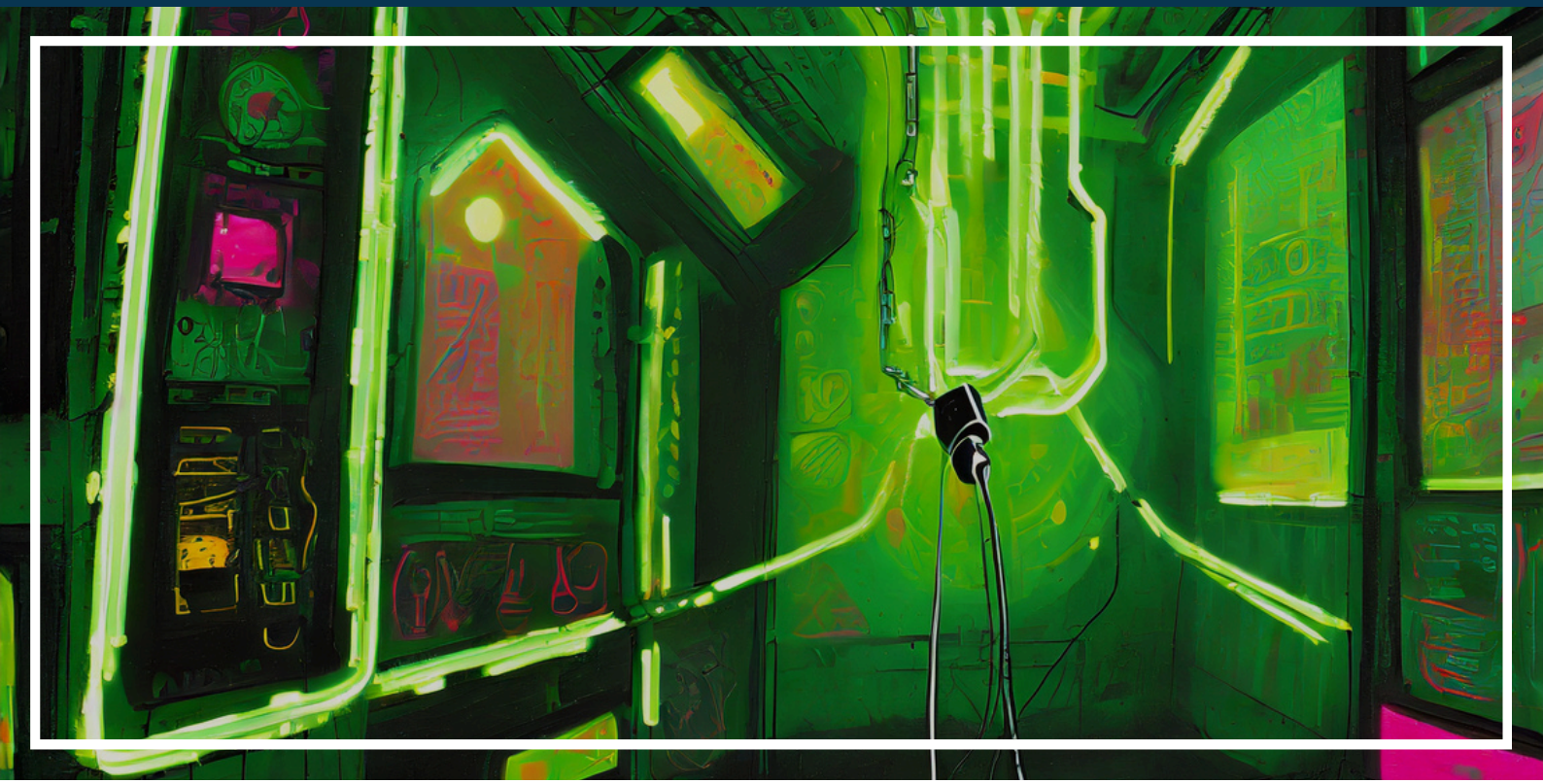

No basta con detener el desarrollo de la IA. Tenemos que apagarlo todo

doi: 10.52749/fh.v4i1.1

ELIEZER YUDKOWSKY

Eliezer Yudkowsky, especialista en inteligencia artificial, comparte sus preocupaciones y propuestas con respecto al futuro impacto de la inteligencia artificial. Fundador del portal LessWrong y cofundador del Machine Intelligence Research Institute (MIRI).

Cómo citar este artículo:

Yudkowsky, E. (2023). No basta con detener el desarrollo de la IA. Tenemos que apagarlo todo (F. López de Pomar, Trad.). *Futuro Hoy, 4*(1), 7-10 https://doi.org/10.52749/fh.v4i1.1. (Trabajo originalmente publicado en 2008).

Una arta abierta publicada el 29 de marzo solicita "que todos los laboratorios de inteligencia artificial detengan de inmediato, durante al menos 6 meses, el entrenamiento de sistemas de inteligencia artificial más poderosos que GPT-4".

Este período de moratoria de 6 meses sería mejor que no tener ninguna moratoria. Tengo respeto por todos los que se han pronunciado y la han firmado. Es un avance marginal. Me abstuve de firmarla porque creo que la carta subestima la gravedad de la situación y pide muy poco para resolverla. El problema clave no es la inteligencia "competitiva con los humanos" (como lo expresa la carta abierta); es lo que sucede después de que la IA supere la inteligencia humana. Los umbrales clave pueden no ser obvios, definitivamente no podemos calcular de antemano lo que sucederá y actualmente parece imaginable que un laboratorio de investigación cruce líneas críticas sin darse cuenta.

Muchos investigadores inmersos en estos temas, incluido yo mismo, esperan que el resultado más probable de construir una IA superinteligente, bajo circunstancias remotamente similares a las actuales, sea que literalmente todos en la Tierra mueran. No es un "tal vez posiblemente alguna remota posibilidad", sino como un "eso es lo obvio que sucedería". No se trata de que no puedas, en principio, sobrevivir creando algo mucho más inteligente que tú; es que requeriría precisión, preparación y nuevos conocimientos científicos, y probablemente no tener sistemas de IA compuestos por matrices inescrutables de números fraccionarios gigantes.

Sin esa precisión y preparación, el resultado más probable es una IA que no haga lo que queremos y que no se preocupe por nosotros ni por la vida consciente en general. Ese tipo de cuidado es algo que en principio se podría infundir en una IA, pero no estamos preparados y actualmente no sabemos cómo hacerlo. En ausencia de ese cuidado, obtenemos "la IA no te ama, ni te odia, y estás hecho de átomos que puede usar para otra cosa".

El resultado probable de que la humanidad se enfrente a una inteligencia superhumana opuesta es una pérdida total. Metáforas válidas incluyen "un niño de 10 años tratando de jugar al ajedrez contra Stockfish 15", "el siglo XI tratando de luchar contra el siglo XXI" y "Australopithecus tratando de luchar contra Homo sapiens".

Para visualizar una IA superhumana hostil, no imagines a un pensador sin vida con conocimientos enciclopédicos que reside en Internet y envía correos electrónicos malintencionados. Visualiza una civilización alienígena completa, pensando a velocidades millones de veces mayores que las humanas, inicialmente confinada a computadoras, en un mundo de criaturas que son, desde su perspectiva, muy estúpidas y muy lentas. Una IA lo suficientemente inteligente no se mantendrá confinada a las computadoras durante mucho tiempo. En el mundo de hoy, puedes enviar secuencias de ADN por correo electrónico a laboratorios que producirán proteínas a pedido, lo que permite que una IA inicialmente confinada a Internet construya formas de vida artificial o pasar directamente a la fabricación molecular postbiológica.

Si alguien construye una IA demasiado poderosa, en las condiciones actuales, espero que cada miembro de la especie humana y toda la vida biológica en la Tierra mueran poco después.

No hay un plan propuesto sobre cómo podríamos hacer algo así y sobrevivir. La intención declarada abiertamente de OpenAI es hacer que alguna futura IA resuelva nuestros problemas de alineación de IA. Solo escuchar que este es el plan debería ser suficiente para hacer que cualquier persona sensata entre en pánico. El otro laboratorio líder en IA, DeepMind, no tiene ningún plan en absoluto.

Una nota aparte: ninguno de estos peligros depende de si las IA son o pueden ser conscientes; es intrínseco a la noción de sistemas cognitivos poderosos que optimizan de manera rigurosa y calculan resultados con criterios suficientemente complicados. Dicho esto, sería negligente en mis deberes morales como ser humano si no mencionara también que no tenemos idea de cómo determinar si los sistemas de IA son conscientes de sí mismos, ya que no tenemos idea de cómo decodificar lo que sucede en esas gigantescas matrices inescrutables, y por lo tanto podríamos crear de manera inadvertida mentes digitales que sean verdaderamente conscientes y que deberían tener derechos y no deberían ser propiedad de nadie.

La regla que la mayoría de las personas conscientes de estos problemas habría respaldado hace 50 años era que si un sistema de IA puede hablar con fluidez y dice que es consciente de sí mismo y exige derechos humanos, eso debería ser una parada definitiva para que las personas lo posean y lo utilicen más allá de ese punto. Ya hemos superado esa antigua línea en la arena. Y eso

probablemente sea correcto; estoy de acuerdo en que las IA actuales probablemente solo están imitando el habla de la autoconciencia a partir de sus datos de entrenamiento. Pero señalo que, con la poca comprensión que tenemos de los entresijos de estos sistemas, en realidad no sabemos.

Si ese es nuestro estado de ignorancia para GPT-4, y GPT-5 es el mismo salto de capacidad gigantesco que de GPT-3 a GPT-4, creo que ya no podremos decir justificadamente "probablemente no es consciente" si permitimos que las personas creen GPT-5. Será simplemente "no lo sé; nadie lo sabe". Si no puedes estar seguro de si estás creando una IA consciente de sí misma, esto es alarmante no solo por las implicaciones morales de la parte "consciente de sí misma", sino porque la incertidumbre significa que no tienes idea de lo que estás haciendo y eso es peligroso y debes detenerte. El 7 de febrero, Satya Nadella, CEO de Microsoft, se jactó públicamente de que el nuevo Bing haría que Google "salga y demuestre que pueden bailar". "Quiero que la gente sepa que los hicimos bailar", dijo.

Así no habla el CEO de Microsoft en un mundo cuerdo. Muestra una brecha abrumadora entre lo seriamente que estamos tomando el problema y lo seriamente que deberíamos haberlo tomado hace 30 años.

No vamos a cerrar esa brecha en seis meses.

Se necesitaron más de 60 años desde que se propuso y se estudió por primera vez la noción de Inteligencia Artificial, hasta que alcanzamos las capacidades actuales. Resolver la seguridad de la inteligencia superhumana, no una seguridad perfecta, sino la seguridad en el sentido de "no matar literalmente a todos", podría tomar al menos la mitad de ese tiempo de manera muy razonable. Y lo que sucede al intentarlo con una inteligencia superhumana es que, si te equivocas en el primer intento, no puedes aprender de tus errores, porque estarás muerto. La humanidad no aprende del error y se levanta y vuelve a intentarlo, como en otros desafíos que hemos superado en nuestra historia, porque todos habremos desaparecido. Intentar hacer algo correctamente en el primer intento realmente crítico es una solicitud extraordinaria, tanto en la ciencia como en la ingeniería. No estamos aplicando ni de cerca el enfoque requerido para hacerlo con éxito. Si consideráramos cualquier cosa en el campo naciente de la Inteligencia Artificial General con los estándares de rigor ingenieril que se aplican a un puente diseñado para llevar unos cuantos miles de autos, todo el campo se cerraría mañana.

No estamos preparados. No estamos en camino de estar significativamente más preparados en un futuro previsible. No hay un plan. El progreso en las capacidades de la IA está avanzando mucho, mucho más rápido que el progreso en la alineación de la IA o incluso el progreso en la comprensión de qué diablos está sucediendo dentro de esos sistemas. Si realmente seguimos adelante, todos moriremos.

Muchos investigadores que trabajan en estos sistemas piensan que nos dirigimos hacia una catástrofe, y muchos de ellos se atreven a decirlo en privado que en público; pero piensan que no pueden detener por sí solos el avance continuo, que otros seguirán incluso si ellos renuncian a sus trabajos. Y así, todos piensan que podrían seguir adelante. Esta es una situación estúpida y una forma indigna de que la Tierra muera, y el resto de la humanidad debería intervenir en este punto y ayudar a la industria a resolver su problema de acción colectiva.

Algunos de mis amigos me han informado recientemente que cuando las personas fuera de la industria de la IA escuchan por primera vez sobre el riesgo de extinción debido a la Inteligencia Artificial General, su reacción es "tal vez no deberíamos construir AGI, entonces".

Escuchar esto me dio un pequeño destello de esperanza, porque es una reacción más simple, más sensata y francamente más cuerda de lo que he estado escuchando en los últimos 20 años al intentar que alguien en la industria tome las cosas en serio. Cualquier persona que hable de manera tan sensata merece escuchar lo grave que es realmente la situación, y no que se le diga que una moratoria de seis meses lo solucionará.

El 16 de marzo, mi pareja me envió este correo electrónico. (Más tarde me dio permiso para citarlo aquí).

"Nina perdió un diente. De la manera habitual en que los niños lo hacen, no por descuido. Ver cómo GPT-4 supera esas pruebas estandarizadas el mismo día en que Nina alcanza un hito de su infancia generó una oleada emocional que me dejó sin aliento por un momento. Todo está avanzando demasiado rápido. Me preocupa que compartir esto aumente tu propio dolor, pero prefiero que lo sepas a que cada uno de nosotros sufra en soledad".

Cuando la conversación interna se centra en el dolor de ver a tu hija perder su primer diente y pensar que no tendrá la oportunidad de crecer, creo que hemos superado el punto de jugar ajedrez político sobre una moratoria de seis meses.

Si hubiera un plan para que la Tierra sobreviva, aun-

que solo sea pasando una moratoria de seis meses, respaldaría ese plan. Pero no hay tal plan.

Esto es lo que realmente se debe hacer:

La moratoria sobre nuevas ejecuciones de entrenamiento a gran escala debe ser indefinida y mundial. No puede haber excepciones, incluyendo gobiernos o militares. Si la política comienza con Estados Unidos, entonces China debe ver que Estados Unidos no busca una ventaja, sino que está tratando de prevenir una tecnología horrendamente peligrosa que no puede tener un verdadero propietario y que matará a todos en Estados Unidos, en China y en la Tierra. Si tuviera libertad infinita para escribir leyes, podría hacer una única excepción para las IA entrenadas exclusivamente para resolver problemas en biología y biotecnología, no entrenadas con texto de Internet y no hasta el punto en que comiencen a hablar o planificar; pero si eso complicara remotamente el problema, inmediatamente abandonaría esa propuesta y diría que simplemente lo apaguemos todo.

Cierren todos los grandes grupos de GPU (las grandes granjas de computadoras donde se refinan las IA más poderosas). Detengan todas las ejecuciones de entrenamiento a gran escala. Establezcan un límite de cuánta potencia informática se permite utilizar en el entrenamiento de un sistema de IA y reduzcan ese límite en los próximos años para compensar algoritmos de entrenamiento más eficientes. Sin excepciones para gobiernos y militares. Lleguen a acuerdos multinacionales inmediatos para evitar que las actividades prohibidas se trasladen a otros lugares. Rastreen todas las GPU vendidas. Si la inteligencia indica que un país fuera del acuerdo está construyendo un grupo de GPU, tengan menos miedo de un conflicto armado entre naciones que de que se viole la moratoria; estén dispuestos a destruir un centro de datos rebelde mediante un ataque aéreo.

No presenten nada como un conflicto entre intereses nacionales, dejen claro que cualquiera que hable de carreras armamentistas es un tonto. Que todos vivimos o morimos como uno solo en esto no es una política, sino un hecho de la naturaleza. Háganlo explícito en la diplomacia internacional que prevenir escenarios de extinción por IA se considera una prioridad por encima de prevenir un intercambio nuclear completo, y que los países nucleares aliados están dispuestos a correr cierto riesgo de intercambio nuclear si eso es lo que se necesita para reducir el riesgo de ejecuciones de entrenamiento de IA a gran escala.

Ese es el tipo de cambio de política que haría que mi pareja y yo nos abrazáramos y nos dijéramos que ha ocurrido un milagro y ahora hay una posibilidad de que tal vez Nina pueda vivir. Las personas sensatas que escuchan esto por primera vez y dicen de manera sensata "tal vez no deberíamos hacerlo" merecen escuchar, honestamente, lo que se necesitaría para que eso suceda. Y cuando tu solicitud de política es tan grande, la única forma de que se lleve a cabo es si los responsables de tomar decisiones se dan cuenta de que, si continúan con los negocios como de costumbre y hacen lo políticamente fácil, eso significa que sus propios hijos también morirán.

Apáguenlo todo.

No estamos preparados. No estamos en camino de estar significativamente más preparados en un futuro previsible. No hay un plan. El progreso en las capacidades de la IA avanza mucho, mucho más rápido que el progreso en la alineación de la IA o incluso el progreso en comprender qué diablos está sucediendo dentro de esos sistemas. Si realmente seguimos adelante, todos moriremos, incluidos los niños que no eligieron esto y no hicieron nada malo.

Apáguenlo todo.