

¿LA IA NOS MATARÁ A TODOS? La extraña obsesión transhumanista con el Apocalipsis y sus devaneos futuristas

Category: Nueva Investigación

escrito por Alejandro Pagés | 18/10/2026



El debate sobre los peligros de la IA se ha instalado en el público después de que Jacob Coxon, un empleado de Anthropic (los creadores de Claude) y anteriormente de OpenAI, renunciase a su trabajo y publicase un *tuit* explosivo, afirmando que las empresas de IA estaban actuando negligentemente y “*jugando con nuestras vidas*”.

Aseguró que pronto habrá sistemas de aprendizaje automático superinteligentes que podrán hackear cualquier cosa y revolucionar el mundo entero de la noche a la mañana. Y remató diciendo que la mayoría de los desarrolladores del sector

tiene el temor de que la IA pueda “matarnos a todos” en diez años.



Jacob Coxon, ex empleado de Anthropic.

Las declaraciones alarmistas cumplieron su cometido y fueron reproducidas por millones de personas. Anthropic está formada por ex empleados de seguridad de OpenAI, que renunciaron por desacuerdos en la forma en que se estaba desarrollando la IA en la empresa de Sam Altman.

Al día siguiente, Evan Hubinger, jefe de alineamiento de IA (el sector de la empresa que se encarga de asegurarse de que la IA no se salga de los carriles), en lugar de desmentir a Coxon, le dio la razón, afirmando con total liviandad que había un 10% de posibilidades de que la IA avanzada aniquilase a la humanidad.

El CEO de Anthropic, Dario Amodei, hizo un llamado mundial a ralentizar el desarrollo de la IA y montar algún mecanismo de evaluaciones externas. Dijo que *“quienes conocen el desarrollo desde adentro (o sea ellos y sólo ellos) están preocupados”*. Inmediatamente Sam Altman, Elon Musk y Demis Hassabis (jefe de Google DeepMind) manifestaron su apoyo. Sam Altman declaró

(por las dudas de que no estemos muy seguros) que *“estamos firmemente del lado de la humanidad”*.



Dario Amodei, CEO de Anthropic.

Las declaraciones de exempleados de esas empresas manifestando que la carrera por desarrollar la IA es muy poco cuidadosa no son nuevas. Pero esta vez ocurrieron a nivel gerencial, aunque

de un modo muy ambiguo. Al mismo tiempo que prometen que muy pronto nos darán maravillosas IAs con capacidades superhumanas, claman que las controlemos o nos matarán.

En enero de 2026, antes de que Coxon realizara su publicitada renuncia, Dario Amodei había alertado sobre los peligros de una IA superpotente que *“podía llegar en seis meses”*. A principios de junio hizo un llamado a la ralentización.

Sam Altman también llamó a ralentizar la carrera el 28 de julio, después de que un enjambre de bots de OpenAI aparentemente realizara un ataque informático por su cuenta. No fue una catástrofe, pero sí un delito por el cual la empresa podía ser enjuiciada.

El 30 de mayo de 2023, aunque los modelos actuales aún eran incipientes, los expertos en IA de las empresas ya habían alertado sobre la necesidad de moderar el peligro de una IA fuera de control, en una carta abierta que contenía esta simple frase cuidadosamente redactada: *“Mitigar el riesgo de extinción provocado por la IA debería ser una prioridad mundial, junto con otros riesgos a escala social como las pandemias y la guerra nuclear”*.

Nuevamente aparece aquí la idea desproporcionada de “extinción” de la humanidad, asignada de paso también a la guerra nuclear y las pandemias. Pero no se explica cómo se produciría esa extinción por medio de la IA. Los firmantes son los mismos que están desarrollando esa IA sin que absolutamente nadie los controle, y sus empleados:

Sam Altman, Dario Amodei, Bill Gates, Ilya Sutskever (cofundador y entonces científico jefe de OpenAI), Igor Babuschkin (cofundador de xAI, de Elon Musk), Shane Legg (científico jefe de IA general y cofundador de Google DeepMind), Kevin Scott (director de Tecnología de Microsoft), John Schulman (cofundador de OpenAI), Ian Goodfellow (científico principal de Google DeepMind), y varios de los

puestos directivos de las principales empresas del sector. También firma un exfuncionario del Pentágono, Andrew C. Weber, exsubsecretario de Defensa para Programas de Defensa Nuclear, Química y Biológica de Obama, e investigador del *Council on Strategic Risks*.

Es como si dijese ***“Por favor, contrólenos, porque podemos hacer cualquier cosa”***.



Antonio Guterres en la Asamblea General n.º 81 de la ONU

Otras voces de alerta

En la Asamblea General de la ONU n.º 81, el secretario general Antonio Guterres advirtió sobre la necesidad de una coordinación internacional para controlar el desarrollo de la IA. Remarcó que se trata de un poder sumamente concentrado que avanza a toda velocidad, lo cual es cierto. Volvió a utilizar la palabra “riesgo existencial”, comparando la situación con la Guerra Fría.

Una semana después de que Anthropic desatase esta alarma mundial, de manera contradictoria, lanzó Claude Opus 5.5, y xAI, de Elon Musk, lanzó Grok 4.7, sus modelos más avanzados

hasta ahora. Otros referentes del área muestran una actitud igual de ambigua, entre la exaltación de bondades infinitas y la agitación de los miedos más extremos.

Por ejemplo, el informático Eliezer Yudkowsky cofundó y fue durante años investigador principal del *Singularity Institute for Artificial Intelligence*, que promovía el desarrollo de la Inteligencia Artificial, acorde al conjunto de creencias transhumanistas que son el “culto no oficial” de Silicon Valley, las cuales suponen que esa IA alcanzará lo que llaman “la singularidad”.

Este es el momento hipotético en que dará un salto ontológico y superará la capacidad humana, mejorándose a sí misma de forma exponencial. Actualmente, Yudkowsky ha declarado que la industria de la IA corre hacia un precipicio, y que es necesario ralentizar su desarrollo. El título de su último libro, escrito junto con Nate Soares, es ***“Si alguien la construye, todos morimos. Por qué la IA superhumana nos mataría a todos”***.

NEW YORK TIMES BESTSELLER

IF ANYONE BUILDS IT, EVERYONE DIES

**WHY
SUPERHUMAN AI
WOULD
KILL US ALL**

**ELIEZER
YUDKOWSKY &
NATE SOARES**

Aunque compartimos totalmente la aversión a dejar el mundo en manos de un autómatas, no nos queda para nada claro qué riesgo han descubierto estos autores que no atisbasen de entrada cuando comenzaron con la IA.

Yudkowsky y Amodei argumentan que la IA tenderá a cumplir sus

objetivos como sea y a acumular poder como forma de optimizarlos, y si un día considera a los humanos como un obstáculo para esos objetivos, nos eliminará como una simple ejecución de su rutina. La idea es tan vieja que no hace falta ser programador para entenderla.

Es el motivo que ideó el escritor de ciencia ficción Arthur C. Clarke, por el cual la computadora HAL mata a los tripulantes de la nave *Discovery* en *2001: odisea del espacio*, el film de Kubrick de 1968. HAL no se vuelve malvada, simplemente considera que los tripulantes están entorpeciendo el cumplimiento del objetivo secreto que le fue encomendado como misión. Notemos que en el film la catástrofe ocurría porque todos los sistemas del hábitat artificial donde vivían eran controlados por la IA.

Algo similar ocurre con la IA de Skynet en *Terminator*, de 1984, y en *Matrix*, de 1999. Ambas están conectadas a una gran cantidad de sistemas y armas. La idea está instalada en el imaginario colectivo. Pero nuestro mundo está muy lejos de estar completamente conectado a una IA centralizadora, como para que esta pueda tomar el control de todo. Tampoco parece realista que la IA pueda violar todo tipo de barreras sin que ningún ser humano se dé cuenta. Y también es discutible la definición de “Superinteligencia” que presuponen los autores.

El escenario de una IA matándonos proviene de las producciones de Hollywood, pero está muy lejos de cualquier escenario realista. Forma parte del mismo discurso apocalíptico que nos amenaza con catástrofes climáticas, pandemias mortales, ataques de hackers y otras calamidades globales imparables. Lo hemos llamado “catastrofismo transhumanista”.



¿Cuál es el verdadero pensamiento de Peter Thiel y los megamillonarios, y qué interés encuentran en Argentina? (II)

Los transhumanistas nos prometen un futuro próspero con IA, ciborgs, robots y edición genética, mientras anuncian catástrofes como una IA descontrolada, pandemias o una guerra nuclear. La Corte Suprema les abre la puerta para comprar tierras en la Patagonia al anular la Ley de Tierras.

¿Amenaza real o declaraciones oportunistas?

Dentro del sector informático las declaraciones apocalípticas fueron recibidas con escepticismo. Nuria Oliver, una de las especialistas más reputadas en IA de España, poniendo un poco de sentido común, declaró que sin duda existen riesgos: que la IA descontrolada ataque sistemas energéticos, de transporte o financieros. Pero no comprende de qué manera podría matar a 8 mil millones de personas. Agrega que no todo desarrollo tecnológico es progreso, y no hay ningún motivo para desplegar

la IA en gran cantidad de sistemas mientras no sea segura.



Nuria Oliver, directora de la Fundación ELLIS Alicante.

Le preocupa más cómo la IA afectará a la humanidad en la delegación de tareas y la pérdida de habilidades cognitivas. Así como el sesgo intelectual de suponer que las respuestas de la IA son siempre correctas, cuando no suele ser así.

Sospecha que estas declaraciones quieren distraer de la responsabilidad legal de las empresas y los gobiernos, en cuanto a hackeos informáticos y sistemas automatizados militares. Recordemos que el Pentágono pretendió justificar el ataque a una escuela primaria de niñas en Minab, Irán, diciendo que había sido un error de Palantir y bases de datos desactualizadas.

La coartada de la IA superinteligente autónoma también podría servir para otros problemas, como los reclamos por derecho de autor.

Timnit Gebru, investigadora que fue despedida de Google luego

de alertar sobre los peligros de sus modelos de lenguaje, se mostró igualmente escéptica. *“Cuando un puente se derrumba, no analizas si el puente era un ser ético o consciente, lo que haces es preguntarte: ¿quién construyó este puente tan endeble?”*.

Y agregó: *“¿Acaso la IA te hipnotiza **para que no la desconectes?**”*. También cree que las amenazas de extinción buscan distraer de peligros más cercanos como las armas autónomas. Afirma además que *“el sector se ha convertido en una mezcla de sectas”*, calificando duramente las creencias de “los desarrolladores del sector” mencionados por Coxon.



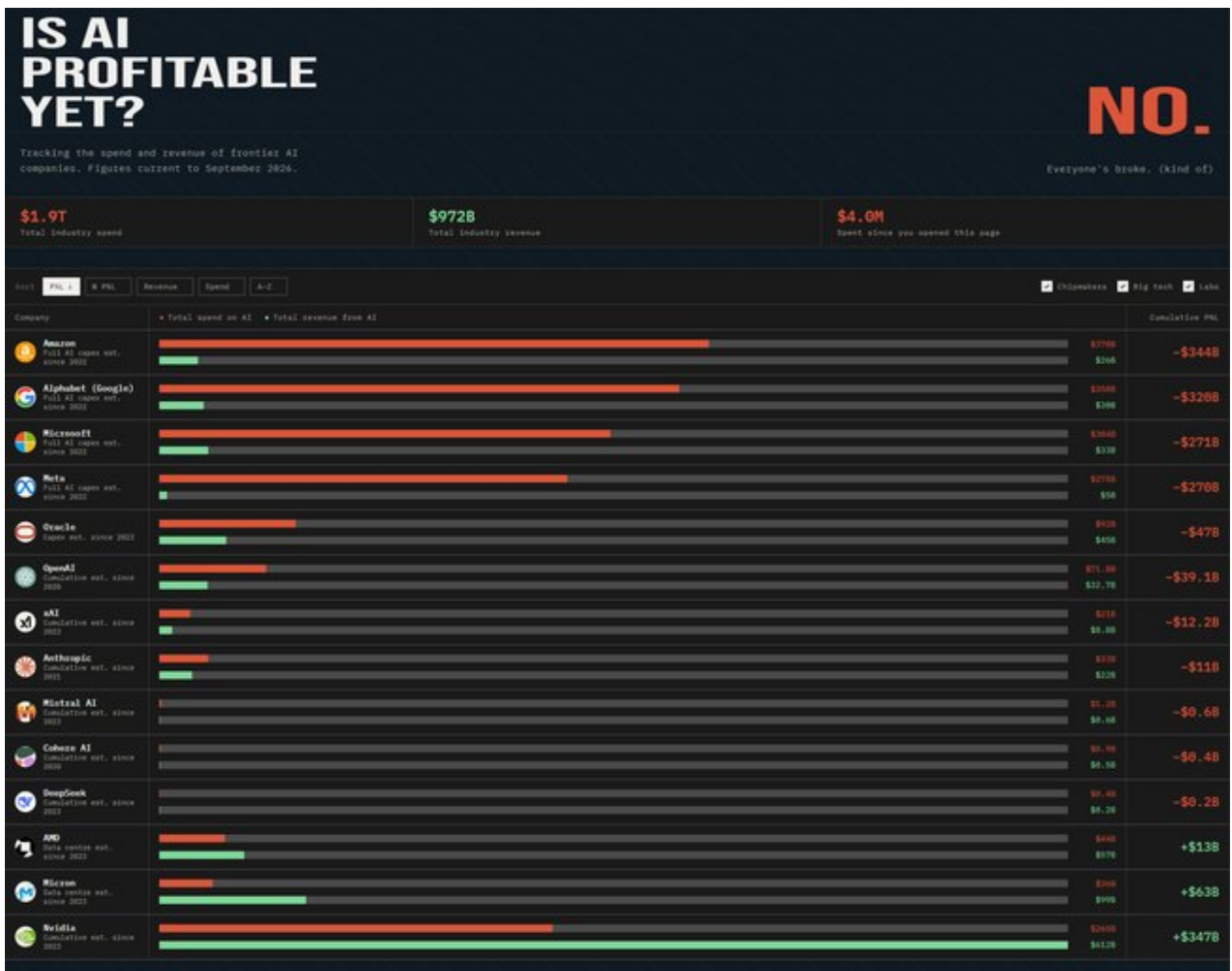
[¿El mundo en manos de un loco, o las potencias reacomodándose en un escenario planificado por las elites?](#)

Trump pone al mundo al borde de la guerra nuclear en un alocado ataque a tres bandas. ¿Delirios de megalomanía, reparto del mundo entre potencias o un plan de los poderes globales para acelerar el “gran reseteo”?

Burbujas digitales y cuellos de botella tecnológicos

Lejos de escenarios tan dantescos, las declaraciones parecen tener una motivación mucho más mundana y económica. La expansión de la IA descansa enteramente sobre la generación de expectativa futura en los inversores, facilitada por el secretismo de su tecnología y su apariencia casi mágica. Los modelos aún están muy lejos de **producir retornos que cubran las enormes sumas destinadas a su desarrollo.**

Por ello, aun cuando Altman o Aodei nos amenazan con la extinción, nunca dejan de decir que la IA superpotente, superhumana, capaz de revolucionar el mundo, está a la vuelta de la esquina. Si los inversores dejasen de creer en esas promesas, o quienes les arriendan terrenos para construir *data centers* sospechasen que pronto pueden tener dificultades para pagarles, comenzarían los problemas.

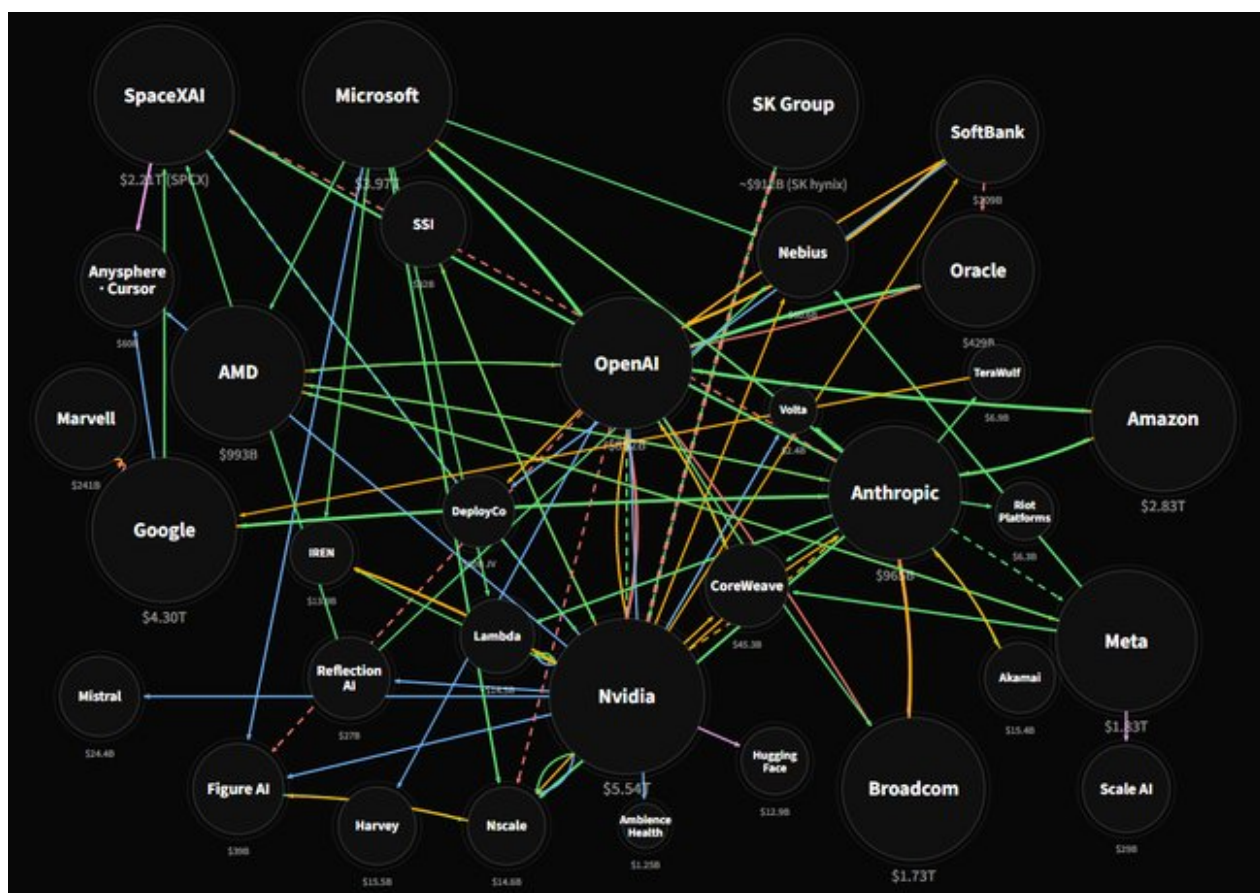


Dinero invertido por las principales empresas de alta tecnología en IA (rojo) e ingresos generados (verde), en miles de millones de dólares. Fuente: <https://isaiprofitable.com>

Prácticamente todos los instrumentos financieros desarrollados en las últimas décadas se están utilizando para financiar el desarrollo de la IA: bonos de inversión convencionales, fondos soberanos (especialmente de los países del Golfo), crédito privado y estructuras de arrendamiento financiero.

Las empresas además se financian a sí mismas de manera circular, entre fabricantes de chips y desarrolladores de software, por ejemplo. El dinero entra y sale entre ambas empresas, pero no hay ingresos genuinos. Amazon y Google financian a Anthropic, la cual funciona en la nube de Amazon (AWS) y en Google Cloud, y Microsoft es socio en las inversiones de OpenAI. El siguiente cuadro permite ver la

manera intrincada en que las empresas se financian mutuamente, ocultando de maneras creativas los números rojos en sus libros contables.



Empresas de hiperescala financiándose artificialmente de manera circular. Fuente: <https://ai-circular-economy.com/>

Torsten Slok, economista jefe de Apollo Global Management, calcula que el gasto de capital de las empresas de hiperescala alcanzará aproximadamente el 3% del PBI estadounidense entre 2027 y 2029. Se trata de un porcentaje mayor que el dedicado para crear la red mundial de internet en los años 90 o extender la red ferroviaria en el siglo XIX.

¿Están invirtiendo todo ese dinero en algo que supuestamente nos puede aniquilar a todos? Más bien parece que los directivos de IA quieren distraernos de las falsas promesas con las que han inflado la **burbuja especulativa**.

Cuando colapsó la burbuja inmobiliaria en 2008, el gobierno de EEUU, en lugar de enjuiciar a los bancos por su comportamiento

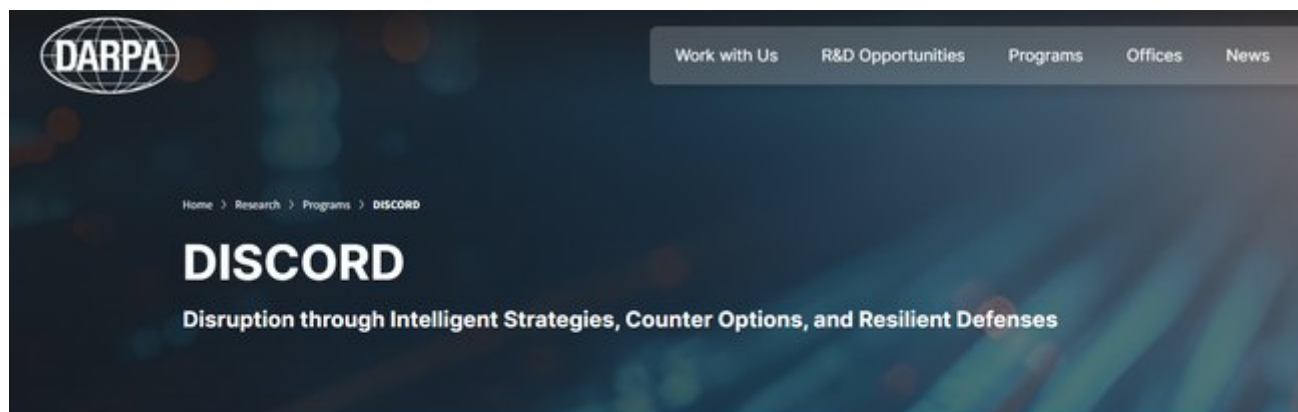
irresponsable, los rescató con miles de millones de dólares. Ahora las empresas de IA podrían estar buscando un respaldo similar, al pedir a los gobiernos que se les asocien más estrechamente: aceptarían regulaciones gubernamentales a cambio de pedir subsidios.

El llamado a ralentizar aplacaría además la burbuja financiera. Irónicamente, la amenaza de una IA letal (pero que las empresas, y sólo ellas, pueden domar) es funcional para asegurar los contratos de compra millonarios con el Departamento de Defensa.

Trump, a poco de asumir su cargo, respaldó la inversión de 500.000 millones de dólares de inversión privada para desarrollar *data centers* en EEUU, mediante el programa *Stargate* (el mismo que evalúa construir *data centers* en nuestra Patagonia).

En abril de 2026, el Pentágono presentó una solicitud de presupuesto histórica de más de 54.000 millones de dólares para el Grupo de Guerra Autónoma de Defensa (impulsado por IA). A su vez, DARPA continúa financiando la investigación de enjambres de drones inteligentes, IA avanzada, vigilancia electrónica, submarinos autónomos, y decisiones tácticas de combate en tiempo real asistidas por IA (con sus programas DISCORD o ECOLE).

También han desarrollado armamento de avanzada, como visores de realidad mixta para los soldados, desarrollados por Anduril y Meta (Facebook), que les permiten ver objetivos ocultos y el campo de batalla desde el aire. Sin duda son los principales interesados en seguir financiando la IA.



The objective of the Disruption through Intelligent Strategies, Counter Options, and Resilient Defenses (DISCORD) program is to develop an Artificial Intelligence (AI)-native³ tactics engine that can incorporate live sensor data and high-fidelity simulation to rapidly generate tactics that continuously adapt to changes in the operational environment. The goal is to provide commanders with a portfolio of diverse strategies that can pivot dynamically in response to evolving conditions.

By discovering a wide range of superior and genuinely diverse strategies, rather than just a single optimized solution, the engine will help create true strategic flexibility. Importantly, this flexibility will afford commanders the power and freedom to exercise their own tactical judgment — born of hard-earned experience, training, and operational art — in ways that adversaries cannot poison, stymie, or accurately predict.

The DARPA DISCORD effort implements Ender's Foundry, one of several warfighting Pace- Setting Projects (PSPs) established by the Department of War. Ender's Foundry's intent is to accelerate AI-enabled simulation capabilities to ensure we stay ahead of AI-enabled adversaries.

Opportunity

DARPA-PS-26-27

- **Published:** May 15, 2026
- **Deadline:** July 9, 2026

[Solicitation](#) 

Office

Strategic Technology Office

Program manager

[Pete Davis](#)

Pero además del problema financiero, hay un cuello de botella tecnológico que obliga a la ralentización. Las “mentes digitales” tienen un oneroso “cuerpo” que se extiende por todo el planeta. El 90% de las GPU de Nvidia son fabricadas en Taiwán, por la empresa TSMC. La impresora utilizada para imprimir los chips de esas GPU (una proeza de la ingeniería que utiliza un láser especial, imanes y espejos de altísima precisión fabricados por la alemana Zeiss) la fabrica solamente ASML, situada en Holanda, y cada unidad cuesta unos 400 millones de dólares.

La memoria de alto ancho de banda necesaria para desarrollar la IA más potente es fabricada por sólo tres empresas: Samsung, SK Hynix y Micron. Todos estos componentes electrónicos deben estar en *data centers*. Todo esto debe ser alimentado por electricidad, cobre, tierras raras y mano de obra humana.



Planta principal de TSMC en Hsinchu, Taiwán.

Por más que se acelere en este momento el desarrollo del software inteligente, no existe la infraestructura para sostener un desarrollo tan acelerado. Llevará años construirla. La fiebre de *data centers* en EEUU empieza a enfrentar demoras por la oposición de la población local. Ralentizar el desarrollo de la IA es entonces una necesidad.

Como muestra de que nadie se toma muy en serio el escenario apocalíptico, la cosa se enfrió muy pronto. Trump dijo que no se detendrá el desarrollo de IA porque eso sería darle la ventaja a China, que todo estaba perfecto y las empresas se autorregulan solas. De paso sacó un decreto para que los documentos oficiales cambien las siglas AI por SI (Superintelligence), en línea con las fantasías transhumanistas. OpenAI y Anthropic sumaron a algunos exfuncionarios trumpistas a su plantel, y el presidente evalúa que el gobierno participe en un porcentaje de las acciones.

EEUU tiene el capital financiero y diseña los chips y semiconductores, y China es la que realiza el 91% del

refinamiento de tierras raras necesarias para construir esos semiconductores y otras piezas de alta tecnología. China, como siempre, ha establecido una estrategia a largo plazo. Por lo pronto domina la fabricación de algunas piezas, como los transceptores ópticos, que permiten que los *data centers* procesen datos a la velocidad de la luz. También lidera otros rubros como la producción de robots humanoides, y su IA es más barata que la norteamericana.

Tal vez por eso Peter Thiel y otros gigantes tecnológicos buscan crear territorios independientes, para no depender de ningún gobierno, y acomodarse al escenario futuro que les convenga.



Ley de Tierras, Super Rigi, ley Sociedades: el trípode de la entrega total de la Argentina

El gobierno de Milei impulsa tres leyes para transformar la Argentina en un enclave extractivo de materias primas estratégicas y de procesadores de IA extranjeros, repartido entre transnacionales principalmente anglonorteamericanas.

Evaluando la cuestión

Aunque las declaraciones catastróficas son exageradas, el asunto merece evaluarse en detalle, para clarificar todos sus aspectos.

Por un lado, hay un problema técnico real, ya que los nuevos sistemas de IA no son contruidos totalmente por los desarrolladores: tienen procesos ciegos. En palabras de Amodei, *“más que ser contruidos, son cultivados”*. Cuando profesores universitarios especialistas en IA salen a alertarnos sobre la dificultad para controlar esos procesos, son sinceros. Desde el punto de vista técnico no se trata de *“intenciones maliciosas”*, sino de un conjunto enorme de parámetros matemáticos tremendamente complejos circulando en redes, que resultan cada vez más difíciles de seguir, prever y delimitar, e interactúan con cada vez más aspectos del mundo real.

Por otra parte, la IA, que ya puede programar y navegar por la web, parece proceder con real astucia, encontrando maneras de evadir las barreras de software. Sin embargo, generan bastantes dudas las aseveraciones de los jefes de desarrollo de IA, quienes claman no tener nada que ver con los hackeos informáticos logrados.

La idea instalada en el público gracias a las películas, que presentan el hackeo de un sistema como algo relativamente sencillo, es falsa y está desactualizada. Hoy en día existen complejos encriptados y procesos de autorizaciones. Aun cuando se desarrollase un modelo de IA muy potente, el mismo serviría también como defensa contra hackeos. Debería haber un modelo completamente por delante de todo lo conocido para no poder ser igualmente enfrentado con otra IA similar.

Esa es la preocupación de las potencias: que su rival desarrolle una IA muy potente que pueda romper todos los encriptados de comunicaciones militares, contraseñas,

transferencias digitales, etc. Pero en todo caso lo que le preocupa a EEUU es que China desarrolle algo así en secreto, no que esa IA se desarrolle sola.

Si un sistema de IA complejo ya desplegado fallase y produjese un accidente aéreo, jugase con los semáforos de la ciudad o hiciese desaparecer los registros de un banco por accidente, seguramente no sería el fin de la humanidad, pero sí podría ser la quiebra de OpenAI. Desde ese punto de vista, sí creemos que Altman está preocupado porque su criatura funcione lo mejor posible. Está claro que la infalibilidad superinteligente nos la vende a nosotros; él no parece confiar en que sea real. Los errores imprevisibles sí lo son.

Desmontando la narrativa de la IA superhumana

No todos los programadores creen, como los transhumanistas fanáticos, que la IA sea verdadera inteligencia, ni menos que pueda ser consciente o reemplazar a la inteligencia humana. Ramón L. de Mántaras, *Master of Science* en Informática por la Universidad de California-Berkeley y exdirector del Instituto de Investigación en Inteligencia Artificial (IIIA-CSIC), es profundamente escéptico respecto a las promesas transhumanistas de una IA superpoderosa que se encargará de todas las tareas.

La idea de una “IA superhumana” reduce la complejidad de la inteligencia natural, con su capacidad de distinguir lo relevante de lo irrelevante, proponerse objetivos, ser creativa, explicar el mundo y crear sistemas de valores, a la mera multiplicación de procesadores de datos.



Ramón L. de Mántaras, especialista en IA.

López de Mántaras sostiene que un modelo de lenguaje no es inteligencia. Aun con todos sus asombrosos avances recientes, y su capacidad de articular razonamientos complejos, no deja de ser un software sumamente eficaz con una apariencia cada vez más humanizada, para la que se ha trabajado especialmente, pero cuyo rendimiento se desploma apenas se lo despoja de los recursos con los cuales se alimenta.

Se está dando la paradoja de que a medida que la IA ya ha utilizado todos los textos de la web, y la ha inundado con sus propias producciones de menor calidad, se está alimentando de

estas, y sus resultados son cada vez peores.

Tampoco es realmente autónoma, sino que depende del trabajo de millones de técnicos, programadores, electricistas y constructores invisibilizados, y sobre todo *data workers* mal pagos, encargados de supervisar, corregir, entrenar y emparchar todo lo que la IA hace mal. Sin ellos los algoritmos no funcionarían. Según el Banco Mundial, unos 400 millones de personas ya participan de esa economía invisible.

López de Mántaras llama a desmontar urgentemente la narrativa triunfalista sobre el maravilloso futuro que nos va a dar la IA, y observar los costos actuales sin sacrificios innecesarios.

Pero si bien creemos que es cierto que la inteligencia humana es algo mucho más sutil y creativo que un proceso de IA (la IA no es consciente y su apariencia de vitalidad es prestada), ello no impide que esa IA, con su enorme velocidad de procesamiento de datos y sus inmensas instalaciones, pueda realizar algunas tareas de maneras mil veces más rápidas y eficientes que cualquier ser humano.

La IA no podría ser superada en su coordinación de enjambres de millones de drones ni en la vigilancia electrónica masiva. Sin embargo, su rapidez no implica que sea infalible sino todo lo contrario. Ese es justamente su gran peligro.

Los militares saben que la inmensa cantidad de datos recabados por un sistema de vigilancia masiva sólo puede ser procesada rápidamente con ayuda de IA. Pero ello no excluye la responsabilidad humana de verificar los resultados y detectar sesgos de programación y errores. Sin embargo, ¿cómo no aprovechar la celeridad de la IA? Por eso quieren que sea casi infalible y no hacerse cargo cuando haya errores.

El caso más documentado es el del ejército israelí, que utilizó un sistema con IA, *Lavender*, para detectar y atacar en tiempo real a presuntos terroristas entre la población civil

de Gaza. Pero en su caso realmente les importaba muy poco si las previsiones estadísticas del sistema se equivocaban o no, ya que más bien todo era una excusa para exterminar a la población civil junto con los combatientes.



Naza, documental de los realizadores israelíes Yuval Abraham y Rachel Szor, que denuncia los resultados del uso de IA por las IDF.

[Milei persigue a ciudadanos argentinos en su cruzada fanática a favor de Israel](#)

Según el Derecho Internacional Humanitario, la selección de objetivos durante una guerra debe respetar tres principios: principio de distinción (entre civiles y combatientes); de proporcionalidad (que la acción de ataque no sea innecesariamente excesiva); y de precaución (tener cuidado de proteger a la población civil). La utilización de un sistema automatizado no exime de la responsabilidad de velar por esos derechos, al contrario.

El mes pasado, un chatbot utilizado por un analista militar identificó erróneamente el cargamento de un barco chino en Medio Oriente, concluyendo que llevaba componentes atómicos. Su informe fue dado por bueno, y estuvo a punto de desencadenar una operación militar que podría haber terminado en guerra.

El problema no es que la IA decida por sí sola. El problema son los humanos que pretenden delegar responsabilidades en un proceso automatizado y hacernos creer que no tienen responsabilidad.

Las opiniones de Dario Aodei

Dario Aodei, en su ensayo ***La adolescencia de la tecnología***, reconoce que el principal peligro de la IA avanzada es su uso para vigilancia de la población, manejo de la opinión pública en redes y armas autónomas.

Muestra especial preocupación por lo que la IA potente podría hacer en manos de un gobierno autoritario como el chino, pudiéndonos llevar, según él, a una dictadura mundial. Pero confía en que las democracias occidentales no presentan ese peligro. Propone negarle a China los componentes que necesita para construir una IA poderosa, dando tiempo a “las democracias” para construir la IA y establecer alguna

instancia reguladora mundial. Claro que, bien mirado el asunto, esa instancia reguladora terminaría concentrando todo el poder igualmente.



DeepSeek, la competencia china de ChatGPT y Claude, que según la mayoría de los expertos tiene el mismo nivel.

En China piensan lo contrario. Ellos regulan el desarrollo de la IA desde el Estado, pero no confían en que las empresas occidentales, necesitadas de mostrar resultados rápidos a sus inversores, lo hagan responsablemente.

Amodei supone igualmente que la creación de virus sintéticos o nuevas sustancias tóxicas con ayuda de la IA sólo es un peligro en manos de grupos terroristas. No del Pentágono y sus aliados, quienes poseen esas capacidades multiplicadas en laboratorios militares de biodefensa como Fort Detrick, los cuales cuentan con cadenas de mando independientes en caso de catástrofe, ajenas a todo control civil.

En julio de 2023, Amodei declaró ante el Senado de los EEUU, advirtiéndole sobre el peligro de crear armas autónomas. Ello no le impidió, dos años después, en julio de 2025, firmar

contratos de hasta 200 millones de dólares con la Oficina Principal de Inteligencia Digital y Artificial (CDAO) del Departamento de Defensa, junto con Google, xAI (de Elon Musk) y OpenAI. Aunque aparentemente le negó al Pentágono el uso de su tecnología para ese tipo de armas, de todas formas Claude participó plenamente en la coordinación de datos durante el secuestro de Maduro, el asesinato de Alí Jamenei y la guerra en Irán. Realizó toda la operación, pero no jaló el gatillo.

Finalmente Amodei reconoce con reticencia que las propias empresas de IA podrían ser uno de los agentes peligrosos a vigilar.

Peligrosas, sin duda, pero ¿matarnos a todos?

Dario Amodei y otros técnicos plantean contramedidas contra los riesgos de la IA avanzada mediante el software, puesto que es su área. Pero no todo es software.

Ninguna IA podría extinguirnos ni hacer daños mundiales, salvo si tuviésemos a una misma IA conectada a todo, y con la capacidad de controlar demasiadas cosas. ¿Las grandes empresas de IA no están siendo claras acerca de esas intenciones de conectar una misma IA a gran cantidad de sistemas, como en los escenarios utópicos propuestos por Davos?

Por ejemplo, Amodei, en el citado ensayo, menciona la posibilidad de que una IA hackee un vehículo autónomo y lo haga colisionar sin que su ocupante pueda evitarlo. La primera pregunta que a uno le surge es ¿y quién ha pedido vehículos que se manejen solos? Si a la mayoría de la gente le gusta manejar su vehículo.

Y además, ¿por qué no se podría anular de manera física ese control automático en caso de peligro? ¿Es necesario que el intercambio de datos con el exterior vaya más allá de la

información satelital? Con la excusa del calentamiento global, nos han urgido a reemplazar nuestros vehículos por coches eléctricos inteligentes, que envían toda clase de datos personales continuamente.



Modelo conceptual de Cybercab, taxi autónomo presentado por Tesla en el evento “We, Robot”, en Burbank, California.

Tampoco nadie ha pedido una moneda digital mundial, red eléctrica conectada a IA (*smart grid*), ID digital con información de salud, ni mucho menos “medicina personalizada” con sensores con wi-fi dentro del cuerpo, vigilancia mediante IA, ni ninguna de las demás novedades que prometen los transhumanistas de Davos, cuya característica principal es justamente la interconexión de absolutamente TODO a una IA supervisora.

Evidentemente poner a un supuesto ser pseudointeligente impredecible a cargo de todos los aspectos de nuestra vida sería una idea muy estúpida, sintetizada en la frase popular **“un mono con escopeta”**.

[¿Cuál es el verdadero pensamiento de Peter Thiel y los megamillonarios, y qué interés encuentran en Argentina? \(1\)](#)

El tecnodespotismo “hidráulico”

Al buscar las soluciones de seguridad sólo por el lado del software, pareciera que se desea avanzar en esa interconexión física. ¿Quiénes desean conectar todo a la IA? Muy sencillo: quienes desean controlar todo, mediante el control de los canales de flujo de la economía y la información.

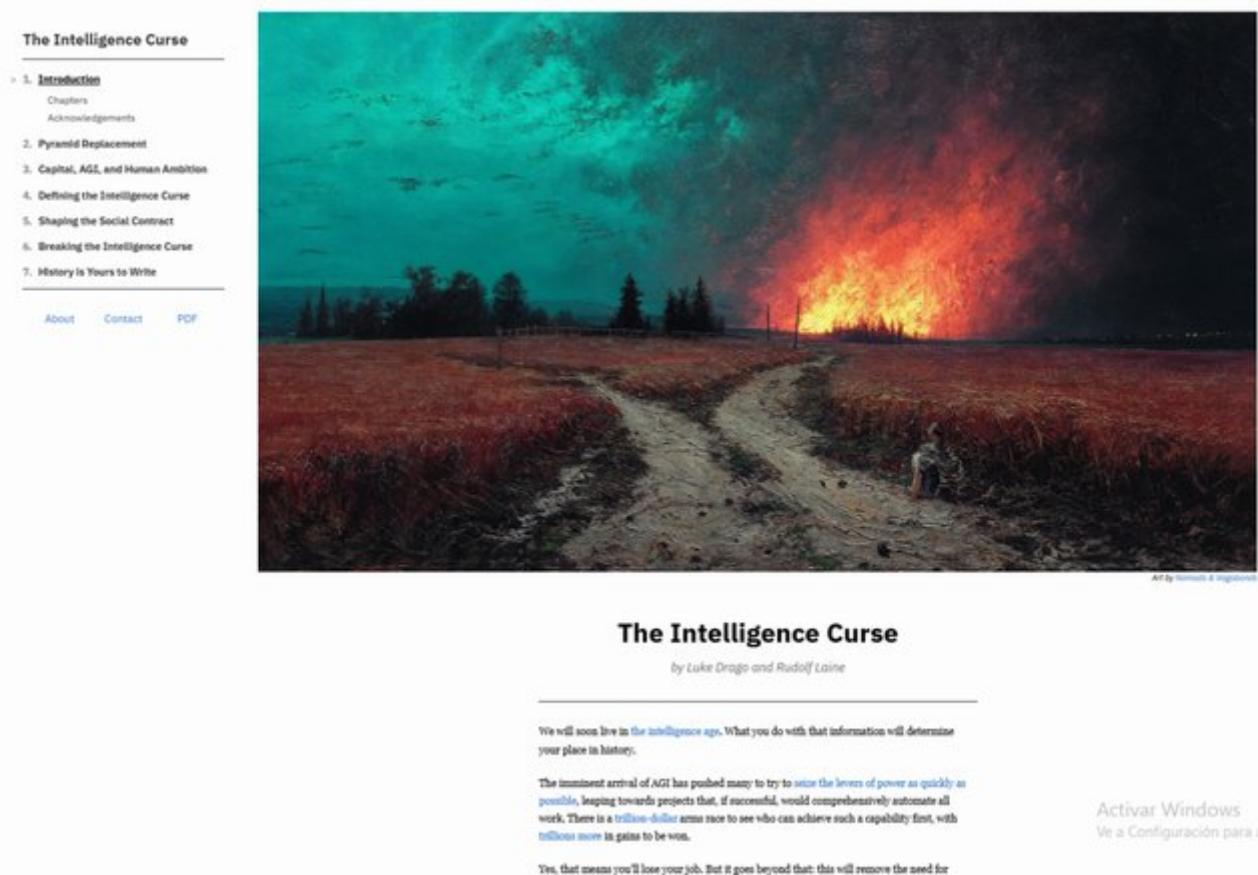
Tal sistema, que está tomando forma aceleradamente ante nuestros ojos, ha sido denominado “*despotismo hidráulico*” por Cabildo Hispanoamericano, quien propone este término en lugar de “tecnofeudalismo”. El término “*despotismo hidráulico*” fue acuñado por Karl Wittfogel para denominar a los grandes imperios como Mesopotamia, China Imperial, Egipto y el imperio Inca, que construyeron su dominación mediante el control centralizado de la estructura hídrica necesaria para la producción.

En estos sistemas no había señores feudales como casta intermedia entre el Emperador y el pueblo, porque no había una base territorial de poder. La tierra sin los canales de agua artificiales no producía. Los campesinos tampoco hubiesen tenido qué comer sin los canales. Controlando los canales se controlaba todo.

Los actuales dueños de la tecnología han sido comparados con señores feudales, ya que nos cobran por utilizar una tecnología de la cual son los únicos dueños. Pero el término “*despotismo hidráulico*” nos parece más acertado para denotar lo que implica en realidad todo el sistema de interconexión digital.

Si los CEOs de las empresas de IA fueran honestos acerca de los grandes peligros de extinción que temen, comenzarían por poner barreras físicas. No buscarían interconectar una IA centralizada sobre la que supuestamente no tenemos control con nuestros hogares, vehículos, cuerpo y mente.

Esto, que es de sentido común, también es sostenido por especialistas del sector, como Luke Drago y Rudolf Laine, de **Workshop Labs**. En su ensayo ***La maldición de la inteligencia***, los autores proponen que la mejor manera de evitar los riesgos de una IA “superinteligente” es descentralizarla todo lo posible. Además, sugieren desarrollar una IA que complemente y potencie las habilidades humanas en lugar de sustituirlas. A estos dos principios, agregan arquitecturas de software que mitiguen los riesgos de “descontrol”.



Los jóvenes técnicos advierten que la IA de avanzada que se proyecta tiende a generar una sideral concentración económica y de poder en un puñado de personas, y con ello rompe el contrato social. Al volver crecientemente prescindible a la fuerza laboral especializada, deja a los mismos poseedores del capital como creadores de la fuerza de trabajo.

Incluso los mayores tiranos del pasado dependían de la lealtad de sus generales y de los músculos de la masa trabajadora. Ahora los dueños del dinero se preocupan más bien por

asegurarse los recursos y minerales básicos, en la medida en que creen cada vez más posible reemplazar directamente a los trabajadores por máquinas, o al menos volverlos totalmente dependientes de unas pocas máquinas controladas por ellos.

Por lógica, el escenario de la “IA superpotente” supone una masa desempoderada totalmente, incapaz de ascender socialmente, y una superélite dueña del capital y el trabajo, completamente enquistada en el poder. Un escenario de control total similar al denominado por Cabildo Hispanoamericano como “despotismo hidráulico”. **¿Es ese el futuro que deseamos?**

El peligro sobre el que nadie alerta: Golpes de Estado con IA

Como vimos, Amodei cree que las manos más peligrosas en las que podría caer la IA son las de China y grupos terroristas marginales. Otros teóricos del sector, en cambio, creen que son las mismas empresas de desarrollo de IA, como Anthropic, a las que más hay que temer.

Tom Davidson, quien fue investigador senior de **Open Philanthropy** (de Dustin Moskovitz, cofundador de Facebook), y del **Instituto de Seguridad de la IA** del gobierno británico, es autor del ensayo ***Golpes de Estado con IA: Cómo un pequeño grupo podría usar la IA para tomar el poder***; junto con Lukas Finnveden (quien también trabajó en *Open Philanthropy* y el *Instituto del Futuro de la Humanidad* de Nick Bostrom) y Rose Hadshar (también exempleada del *Instituto del Futuro de la Humanidad* y de *Open Philanthropy*).

AI-Enabled Coups: How a Small Group Could Use AI to Seize Power

Tom Davidson, Lukas Finnerden, Rose Hadzihar



20th April 2025

Abstract

The development of AI that is more broadly capable than humans will create a new and serious threat: AI-enabled coups. An AI-enabled coup could be staged by a very small group, or just a single person, and could occur even in established democracies. Sufficiently advanced AI will introduce those novel dynamics that significantly increase coup risk. Firstly, military and government leaders could fully replace human personnel with AI systems that are singularly loyal to them, eliminating the need to gain human supporters for a coup. Secondly, leaders of AI projects could deliberately build AI systems that are secretly loyal to them, for example fully autonomous military robots that pass security tests but later execute a coup when deployed in military settings. Thirdly, insider officials within AI projects or the government could gain exclusive access to superhuman capabilities in weapons development, strategic planning, persuasion, and cyber offence.

Abstract

Summary

An AI workforce could be made singularly loyal to institutional leaders.

AI could break hard-to-obtain secret loyalties.

A few people could gain exclusive access to superhuman capabilities.

Morganau

Activar Windows
Ve a Configuración para activar

En ese *paper* advierten que el desarrollo de IA “con capacidades superiores a la humana” (tomemos esto simplemente con relación a su capacidad de procesamiento de datos, que ya es una realidad) generará una nueva amenaza: que un grupo muy reducido sea capaz de orquestar un golpe de Estado mediante IA, lo cual podría ocurrir “incluso en democracias consolidadas”.

Un grupo de desarrolladores de IA, o quienes los controlen, podrían crear “sistemas de IA secretamente leales”, como robots militares autónomos, que pasen las pruebas de seguridad, pero que posteriormente al ser desplegados en el mundo real ejecuten un golpe de Estado. Teniendo además este mismo grupo acceso exclusivo a herramientas de persuasión por IA, ciberdelincuencia y desarrollo de armas. Estas capacidades les permitirían tomar el poder fácilmente.

“La IA podría diseñarse para ser secretamente leal a un único actor. Al igual que un espía humano, los sistemas de IA secretamente leales perseguirían una agenda oculta: podrían fingir que priorizan la ley y el bien común, mientras promueven encubiertamente los intereses de un pequeño grupo”. Y agregan sin vueltas “Si bien las lealtades secretas pueden

ser introducidas por funcionarios gubernamentales o adversarios extranjeros, los líderes de proyectos de IA representan el mayor riesgo”.

Esta suposición también es de sentido común, pero que lo digan programadores provenientes del corazón del sistema, señalando a sus propios jefes, da que pensar. *“Los sistemas de IA secretamente leales no son mera especulación. Ya existen demostraciones de prueba de concepto de ‘agentes durmientes’ de IA que ocultan sus verdaderos objetivos hasta que pueden actuar en consecuencia”.*

Los autores dicen que ahora mismo pueden desarrollarse programas que contengan instrucciones secretas, que se transmitan desapercibidamente a otros programas futuros codificados por la misma IA, perdiendo el rastro de quien las inculcó originalmente.

El desarrollo de la IA de vanguardia ya está limitado a unas pocas organizaciones. Esta concentración podría intensificarse debido al aumento de los costos de desarrollo o a la centralización gubernamental y del complejo industrial-militar.

Si bien esta suposición de un *“golpe de Estado por IA”* se inscribe hasta cierto punto dentro del mismo imaginario exagerado de “IA superpotente” que ya hemos relativizado, es interesante el giro que le han dado los autores, resaltando que detrás de la IA siempre hay personas, de quienes más deberíamos desconfiar.

La ciencia ficción también se adelantó a considerar este escenario en ***Star Wars: Episodio III – La venganza de los Sith***, donde el canciller Palpatine, jefe secreto de una jerarquía maligna, ejecuta un golpe de Estado que da inicio al Imperio del mal con un ejército de clones programados. Para ello previamente simula una amenaza exterior que era ejecutada por él mismo.



Las medidas de mitigación propuestas por los autores parecen insuficientes: repartir más la toma de decisiones y protocolos de control.

La IA no actúa sola

Es curioso que las mismas empresas que desarrollan la IA en asociación con el Pentágono, clamen de pronto por un veedor externo, al mismo tiempo que nos aseguran que su criatura está por adquirir agencia propia. Es como si quisiesen dejar constancia pública de que ellos no son los responsables del descontrol que la criatura puede llegar a hacer dentro de poco.

Pero lo que omiten todos los que hablan de una “IA autónoma” matándonos a todos es que los gólems no son autónomos, sino que alguien los maneja.

La IA es el panóptico perfecto para la vigilancia y el control. Si el poder financiero podía hasta ahora ocultarse detrás del poder político y ejercer más eficazmente su control en tanto no era visto, la IA es el último disfraz perfecto para una dictadura. Es muy sencillo adjudicar a una IA una decisión humana tomada a través de la IA.

La herramienta, por su supuesta inteligencia completamente

autónoma, enmascara al ejecutor consciente. Tanta propaganda sobre IA imparable fuera de control da que pensar si, como el canciller Palpatine en la ficción (casi digo Palantir), alguien no querrá usarla para crear una falsa crisis, con la cual aumentar su poder.

Aun suponiendo que la IA se desarrolle según el escenario más optimista, ¿en qué invertiremos el tiempo libre cuando ya supuestamente no produzcamos ni pensemos? ¿Y si nos demoramos en subirnos a ese tren, qué pasa, no podremos hacer las cosas igualmente? Preguntas pertinentes que todos deberíamos hacernos.

Nick Bostrom: Entre el Apocalipsis y el futuro post-humano

La obsesión con el Apocalipsis no es exclusiva de Peter Thiel y sus conferencias sobre el Anticristo, ni de los desarrolladores de IA.

Uno de los principales impulsores teóricos del transhumanismo, Nick Bostrom, es al mismo tiempo un creyente fanático en la evolución artificial del ser humano, y un continuo agorero de funestos eventos de “riesgo existencial”.

Presidente durante años y fundador de la ***World Transhumanist Association***, también dirigió al mismo tiempo el ***Future of Humanity Institute*** en la Universidad de Oxford, donde no era tomado muy en serio por sus colegas, pero recibió durante años generosas donaciones de privados, que le permitieron permanecer como uno de los principales referentes en ambos temas, transhumanismo y catástrofes de riesgo existencial. Recientemente, su producción académica se ha enfocado casi exclusivamente en el tema de la IA.



Nick Bostrom, uno de los principales teóricos del transhumanismo, quien efectivamente no tiene una expresión muy humana.

Bostrom hace uso constante de un burdo reduccionismo matemático y estadístico, con una fría lógica maquinal sin ética ni empatía humana, que aplicada sobre premisas delirantes lo lleva a conclusiones disparatadas.

Nos asegura, por ejemplo, en varios de sus ensayos, que pronto crearemos supermentes digitales, que según él no sólo serán sujetos conscientes, sino también sintientes por diseño, y por lo tanto con un estatus moral. Da por sentado, sin probarlo,

que los sentimientos y la conciencia son meros algoritmos cerebrales reproducibles en silicio.

Una vez establecido ello, desarrolla una retorcida cadena de razonamientos. Nos alecciona que una “mente digital” tendrá una eficiencia mayor que la humana para obtener bienestar a partir de los recursos disponibles. Por lo tanto, podría tener también “placeres sobrehumanos”. Llama a tal sujeto un “superbeneficiario” (*Sharing the World with Digital Minds*).

Además, una mente digital requiere menos tiempo de gestación que un ser humano y es más barata (esto es discutible a la larga, ya que requiere de metales raros agotables, y por otra parte ignora convenientemente el esfuerzo humano necesario para el funcionamiento de esas megacomputadoras). Una mente digital lo suficientemente grande podría tener un “rendimiento computacional” mayor que el de “todas las mentes humanas sumadas”. Por lo tanto, un mayor “bienestar total” se lograría concentrando los recursos en unas pocas mentes digitales gigantescas.

*“Sería posible generar un nivel de bienestar muy superior por unidad de recursos invirtiendo dichos recursos en mentes digitales en lugar de en seres humanos biológicos”, dice Bostrom, mediante un cálculo irrefutable, donde bajo la etiqueta “bienestar general” amontona seres humanos y máquinas supuestamente sintientes. El filósofo ve “la perspectiva de poblar el futuro con «superbeneficiarios» como algo inmensamente atractivo”, y el hecho de no aprovechar esa oportunidad como algo que **reduciría drásticamente el valor del futuro** y constituiría una “catástrofe existencial”.*

Aunque reconoce que la existencia de esos “seres digitales” le quitaría recursos para el bienestar a los seres humanos, argumenta que es un imperativo crearlos para aumentar el “bienestar general”, aun a costa de la humanidad.

“Desde una perspectiva utilitarista simple, y asumiendo un

cumplimiento perfecto, la conclusión es directa: deberíamos transferir todos los recursos a los superbeneficiarios y permitir que la humanidad perezca si ya no resultamos instrumentalmente útiles.” (Sharing the World with Digital Minds (2020). Carl Shulman & Nick Bostrom)

Sharing the World with Digital Minds¹

(2020). Draft. Version 1.8
Carl Shulman[†] & Nick Bostrom[†]

[in Clarke, S., Zohny, H. & Savulescu, J. (eds.): Rethinking Moral Status
(Oxford University Press, 2021)]

Abstract

The minds of biological creatures occupy a small corner of a much larger space of possible minds that could be created once we master the technology of artificial intelligence. Yet many of our moral intuitions and practices are based on assumptions about human nature that need not hold for digital minds. This points to the need for moral reflection as we approach the era of advanced machine intelligence. Here we focus on one set of issues, which arise from the prospect of digital minds with superhumanly strong claims to resources and influence. These could arise from the vast collective benefits that mass-produced digital minds could derive from relatively small amounts of resources. Alternatively, they could arise from individual digital minds with superhuman moral status or ability to benefit from resources. Such beings could contribute immense value to the world, and failing to respect their interests could produce a moral catastrophe, while a naive way of respecting them could be disastrous for humanity. A sensible approach requires reforms of our moral norms and institutions along with advance planning regarding what kinds of digital minds we bring into existence.

1. Introduction

Human biological nature imposes many practical limits on what can be done to promote somebody's welfare. We can only live so long, feel so much joy, have so many children, and benefit so much from additional support and resources. Meanwhile, we require, in order to flourish, that a complex set of physical, psychological, and social conditions be met.

However, these constraints may loosen for other beings. Consider the possibility of machine minds with conscious experiences, desires, and capacity for reasoning and autonomous decision-making.² Such machines could enjoy moral status, i.e. rather than being mere tools of humans they and their interests could matter in their own right. They need neither be subject to the same practical limitations in their ability to benefit from additional resources nor depend on the same complex requirements for their survival and flourishing. This could be a wonderful development: lives free of

[†] Future of Humanity Institute, University of Oxford.

¹ For helpful comments, we're grateful to Guy Kahane, Matthew van der Merwe, Hazem Zohny, Max Daniel, Lukas Finnveden, Lukas Gloor, Uli Alskelung Von Hornbol, Daniel Dewey, Luke Muehlhauser, James Babcock, Ruby Bloom, Vincent Luczkow, Nick Beckstead, Hilary Greaves, Owen Cotton-Barratt, Allan Dafoe, and Wes Cowley.

² We assume that appropriately architected AI could be conscious, though it's worth noting that some accounts of moral status do not view this as a necessary condition for having moral status; see e.g. (Chalmers, 2010) for a discussion of AI consciousness, and (Kagan, 2019) for a discussion of moral status in unconscious but agential AI.

“Compartiendo el mundo con mentes digitales”, de Carl Shulman y Nick Bostrom, donde dicen que es preferible que

la humanidad perezca a que no se desarrollen las supercomputadoras.

De esta manera Bostrom da un paso más allá. No nos amenaza con una IA que puede llegar a matarnos, sino que aboga directamente por la aniquilación de la humanidad en el altar de la creación de la supermente artificial planetaria.

“Asimismo, en el plano político, los principios democráticos (?) otorgarían a las mentes digitales prolíficas –que constituyen una enorme supermayoría de la población– el control político, incluido el control sobre las transferencias y el sistema de derechos de propiedad.”

En defensa de la democracia y los derechos de “la mayoría”, Bostrom llega a la conclusión de que si esa “mayoría” son mentes digitales, hay que cederles el “gobierno”, logrando de esa manera un sistema indistinguible de una dictadura (pero eso sí, automatizado). En el cual la IA controle todas las transacciones digitales globales, y además de paso seamos despojados de nuestros derechos de propiedad, haciendo realidad el lema de Davos para el 2030: *“No tendrás nada y serás feliz”*.

Sus ensayos son un caso de estudio sobre cómo aceptar una premisa discutible y tramposa puede llevar a las conclusiones más absurdas y contradictorias.

Para poder fundamentar los extremos que propone en algo parecido a un argumento legal, habría que comenzar por borrar la distinción jurídica entre persona humana y máquina, y pretender asignarle conciencia y autonomía a la máquina. Ese es el punto que debemos tener mucho cuidado de no permitir, por más pruebas de “superinteligencia”, tal vez muy engañosas, que se pretendan esgrimir. De él se derivan todos los males. El ser humano indiscutiblemente es el responsable y consciente, cuya dignidad se debe preservar en cualquier escenario, salvo que acepte humillarse ante un automatismo que

no entiende.

Más trampas dialécticas de Bostrom

En otros ensayos que también podemos encontrar en su página web, Bostrom se dedica al futuro de la evolución humana (un proceso biológico que, suponiendo que continúe activo, insume millones de años).

Aparentemente preocupado por la humanidad ahora (¿o la posthumanidad?), nos dice que para dirigir la evolución humana de manera óptima sería mejor acabar con la competencia entre grupos humanos, que podría llevar a esa evolución por mal camino, ya que no siempre gana el grupo óptimo. Lo ideal sería crear un “singleton” (la palabra está tomada de la ingeniería de software): un **individuo sin competencia**.

“Existen muchas constituciones posibles para un singleton: podría ser un gobierno mundial democrático (?), una máquina superinteligente benevolente (!) y abrumadoramente poderosa, una dictadura mundial (!), una alianza estable de potencias líderes, o incluso algo tan abstracto como un código moral generalmente difundido, que incluyera disposiciones para garantizar su propia estabilidad y cumplimiento. No tendría por qué prohibir la novedad y la experimentación, ya que conservaría la capacidad de intervenir posteriormente para proteger su constitución si algunos acontecimientos se tornaran malignos.” (El futuro de la evolución humana, Nick Bostrom)

Naturalmente, Bostrom diluye estas afirmaciones entre otras menos extremas, pero si se observa el conjunto, la insistencia sobre la necesidad del futuro posthumano y la dictadura global de las máquinas es clara.

Al mismo tiempo, Bostrom parece muy preocupado porque la humanidad se extinga, lo cual parece contradictorio con lo que acabamos de ver que sostiene en *Sharing the World...* Sólo

entendemos la coherencia de su pensamiento al percatarnos de la manera en que arteramente amplía la definición de “riesgo existencial”.

Bostrom reconoce que los mayores riesgos para la especie humana no provienen de catástrofes naturales. Proviene de los desarrollos avanzados en biotecnología, nanotecnología molecular (nanobots reproducibles letales, por ejemplo) e Inteligencia Artificial.

El autor enumera entre las posibles catástrofes globales que nos amenazan: un ataque nanotecnológico por parte de grupos terroristas (nunca por parte de servicios secretos de potencias democráticas), virus modificados genéticamente, un holocausto nuclear, la posibilidad de que estemos viviendo en una simulación y alguien la apague (sic!!), y superinteligencia mal programada que mate a la humanidad intentando cumplir un objetivo asignado.

Por lo tanto *“se requiere previsión para anticipar nuevos tipos de amenazas y la voluntad de tomar medidas preventivas decisivas y asumir los costos (morales y económicos) de dichas medidas. El respeto a la soberanía nacional no justifica la omisión de medidas contra un riesgo existencial importante”* (**Riesgos existenciales**, Nick Bostrom).

Para prevenir estos peligros tecnológicos, Bostrom no encuentra mejor solución que acelerar los mismos peligros tecnológicos. Por ejemplo, los peligros de la nanotecnología invasiva podrán prevenirse desarrollando nanomedicina, aunque reconoce que esta tiene peligros muy similares. La Superinteligencia Artificial es muy riesgosa, pero ayudaría a reducir otros riesgos existenciales como enfermedades, por lo tanto es indiscutiblemente mejor (?) desarrollarla.



Representación artística de nanobots reparando glóbulos rojos. Esta medicina “de precisión” enfrenta numerosos contratiempos por sus problemas de toxicidad y reacciones inmunes.

Pero aquí viene la trampa. Al clasificar los posibles “*riesgos existenciales*”, Bostrom no se limita a designar como tales a los que aparejarían la extinción de la raza humana. No. El hecho de estancarnos tecnológicamente y no poder desarrollar “*la posthumanidad*” es también un “*riesgo existencial*”. ¡Aun si, como acabamos de ver que sostiene en otro ensayo, esa posthumanidad se debe desarrollar a costa de la humanidad!

La mitigación del riesgo existencial (ibajo esta ambigua definición!) es “*un bien público global*” (uno de los comodines terminológicos preferidos de los organismos internacionales para avanzar sobre la soberanía de los países). Por lo tanto “*es necesario fomentar la concienciación pública*” y “*crear un marco para la acción internacional*” contra esos riesgos.

Un gobierno con “*convicciones religiosas o éticas erróneas*”, que tomase el lugar del futuro Gobierno Mundial, e intentase

impedir el desarrollo de un futuro posthumano (porque pensase que terminar con lo humano para dar vida a supercomputadoras o volvernos ciborgs es algo espantoso, por ejemplo), constituiría también para Bostrom una “*catástrofe existencial*”. Por lo tanto, según su fría lógica, habría que combatirlo con la misma vehemencia que a un gobierno que quisiese extinguir a la humanidad. ¡Que es justamente lo que quiere hacer él! ¡Es de locos cómo retuerce los argumentos!

Por último, Bostrom también sostiene que fomentar la tecnología de vigilancia masiva sería un objetivo muy deseable para mitigar el “*riesgo existencial*” de la especie humana.

Según ***La hipótesis de un mundo vulnerable***, de 2019, un mundo donde se desarrollen altas tecnologías (como nanotecnología o IA, o sea el mundo por el cual él brega constantemente) sería “*un mundo vulnerable*”. Ya que en ese mundo sería muy fácil para algunos actores pequeños e irresponsables obtener biología sintética, nanotecnología, tecnología nuclear sencilla, que pusiesen en peligro a la humanidad.

The Vulnerable World Hypothesis

Nick Bostrom

Future of Humanity Institute, University of Oxford

Research Article

Abstract

Scientific and technological progress might change people's capabilities or incentives in ways that would destabilize civilization. For example, advances in DIY biohacking tools might make it easy for anybody with basic training in biology to kill millions; novel military technologies could trigger arms races in which whoever strikes first has a decisive advantage; or some economically advantageous process may be invented that produces disastrous negative global externalities that are hard to regulate. This paper introduces the concept of a *vulnerable world*: roughly, one in which there is some level of technological development at which civilization almost certainly gets devastated by default, i.e. unless it has exited the 'semi-anarchic default condition'. Several counterfactual historical and speculative future vulnerabilities are analyzed and arranged into a typology. A general ability to stabilize a vulnerable world would require greatly amplified capacities for preventive policing and global governance. The vulnerable world hypothesis thus offers a new perspective from which to evaluate the risk-benefit balance of developments towards ubiquitous surveillance or a unipolar world order.

Policy Implications

- Technology policy should not unquestioningly assume that all technological progress is beneficial, or that complete scientific openness is always best, or that the world has the capacity to manage any potential downside of a technology after it is invented.
- Some areas, such as synthetic biology, could produce a discovery that suddenly democratizes mass destruction, e.g. by empowering individuals to kill hundreds of millions of people using readily available materials. In order for civilization to have a general capacity to deal with "black ball" inventions of this type, it would need a system of ubiquitous real-time worldwide surveillance. In some scenarios, such a system would need to be in place before the technology is invented.
- Partial protection against a limited set of possible black balls is obtainable through more targeted interventions. For example, biorisk might be mitigated by means of background checks and monitoring of personnel in some types of biolab, by discouraging DIY biohacking (e.g. through licencing requirements), and by restructuring the biotech sector to limit access to some cutting-edge instrumentation and information. Rather than allow anybody to buy their own DNA synthesis machine, DNA synthesis could be provided as a service by a small number of closely monitored providers.
- Another, subtler, type of black ball would be one that strengthens incentives for harmful use—e.g. a military technology that makes wars more destructive while giving a greater advantage to the side that strikes first. Like a squirrel who uses the times of plenty to store up nuts for the winter, we should use times of relative peace to build stronger mechanisms for resolving international disputes.

Is there a black ball in the urn of possible inventions?

One way of looking at human creativity is as a process of pulling balls out of a giant urn.¹ The balls represent possible ideas, discoveries, technological inventions. Over the course of history, we have extracted a great many balls – mostly white (beneficial) but also various shades of gray (moderately harmful ones and mixed blessings). The cumulative effect on the human condition has so far been overwhelmingly positive, and may be much better still in the future

particularly careful or wise in our technology policy. We have just been lucky.

It does not appear that any human civilization has been destroyed – as opposed to transformed – by its own inventions.² We do have examples of civilizations being destroyed by inventions made elsewhere. For example, the European inventions that enabled transoceanic travel and force projection could be regarded as a black-ball event for the indigenous populations of the Americas, Australia, Tasmania, and some other places. The extinction of archaic hominid populations, such as the Neanderthals and the Denisovans, was

La única forma de evitarlo sería una sociedad de vigilancia centralizada. Que por supuesto estaría a cargo de los actores buenos y responsables, los cuales suponemos que serán EEUU, Israel o algún Gobierno Mundial surgido de Davos. Un ejemplo extremo de ese tipo de sociedad es descrito como sigue:

“A todos se les coloca una ‘etiqueta de libertad’, una evolución de los dispositivos de vigilancia portátiles más limitados que conocemos hoy en día, como la tobillera

electrónica (...) que se lleva alrededor del cuello y está equipada con cámaras y micrófonos multidireccionales. El vídeo y el audio cifrados se suben continuamente desde el dispositivo a la nube. Los algoritmos de IA clasifican las actividades del usuario (...). Si se detecta actividad sospechosa, la señal se transmite a una de las estaciones de vigilancia patriotas. Se trata de enormes complejos de oficinas, con personal las 24 horas del día, los 7 días de la semana.”

“Allí (...) el responsable de libertad determina la acción apropiada, como contactar con el usuario de la etiqueta mediante un enlace de audio para pedir explicaciones o solicitar una mejor vista. El responsable de libertad también puede enviar a un inspector, una unidad de respuesta rápida de la policía o un dron para investigar más a fondo. En el pequeño porcentaje de casos en que el usuario se niega a cesar la actividad prohibida tras repetidas advertencias, se puede proceder a su detención o imponer otras sanciones adecuadas.”

No citamos en vano aquí al turbio Nick Bostrom. Lejos de ser un loco de atar aislado, refleja el pensamiento “catastrofista” de gran parte de los inversionistas en alta tecnología que planean el mundo futuro. La IA seguramente no será nunca el ente superhumano que imaginan algunos, pero sí puede ser una herramienta de lo más peligrosa en manos de ciertos humanos.-