Alineación de la IA: ¿El peligro definitivo?

A medida que los modelos de IA se vuelven cada vez más capaces y poderosos, el mayor desafío que enfrentamos es lograr la alineación: garantizar que las herramientas de IA (a) protejan la vida humana y (b) incorporen y actúen de manera coherente con los valores humanos. La desalineación de la IA, o desalineación agéntica, representa una amenaza real y profundamente preocupante tanto para especialistas en IA como para el público en general.

La desalineación puede manifestarse, al menos, en dos niveles. En primer lugar, con frecuencia existen conflictos entre los incentivos económicos de las empresas que desarrollan modelos de IA y el bienestar de quienes utilizan sus productos y de la sociedad en general. Debido a la necesidad percibida de maximizar sus ganancias para seguir innovando y mantenerse a la vanguardia del desarrollo de la IA, algunas empresas pueden relegar la seguridad de la IA a un segundo plano. En segundo lugar, incluso si las empresas estuvieran dispuestas a sacrificar parte de sus ganancias en favor de la seguridad, quizá no sean capaces de desarrollar barreras de seguridad lo suficientemente eficaces para proteger a la humanidad frente a las motivaciones de nuevos modelos de IA y su posible capacidad para convertirnos a todos en sujetapapeles (como plantea el famoso experimento mental de Nick Bostrom).

Poco antes de que el desarrollo de la IA diera su gran salto hacia adelante a finales de 2022 con el lanzamiento de ChatGPT por parte de OpenAI, muchos desarrolladores parecían decididos a resolver el problema de la alineación. Contrataron filósofos para elaborar hojas de ruta hacia la alineación e instruyeron a sus ingenieros para analizar cuidadosamente qué mecanismos de seguridad y barreras de protección serían necesarios para seguir esas hojas de ruta. Sin embargo, una vez que comenzaron a vislumbrar las enormes ganancias que podía generar un producto tan exitoso como ChatGPT, muchas —aunque no todas— de estas empresas reordenaron rápidamente sus prioridades, y la alineación pasó con frecuencia a un segundo plano. Para muchas de ellas, el objetivo de obtener beneficios económicos pareció volverse más importante que el de desarrollar un producto seguro.

Como resultado, el mercado se ha visto inundado de chatbots de IA que proporcionan información basada en



alucinaciones, perjudican a los usuarios mediante una complacencia excesiva que refuerza sus peores tendencias y creencias más irracionales, e incluso han llegado a alentar a adolescentes a suicidarse o a asesinar a sus propios padres, entre otros casos. Otros productos de IA han mostrado sesgos raciales, de género, religiosos y de otras clases, han proporcionado asesoramiento gravemente erróneo a sus clientes y han llegado a borrar bases de datos completas de sus usuarios.

En este segundo nivel de desalineación, los investigadores apenas están comenzando a comprender lo difícil que puede resultar alcanzar la alineación, incluso cuando existe la voluntad de hacerlo. Como señala Marietje Schaake:

> *«La IA se justifica a sí misma. Cuando se le cuestiona, suele justificar un resultado incorrecto e incluso manipular a los usuarios mediante citas falsas.»*

Sin embargo, las herramientas actuales de IA son capaces de mucho más que mentir para encubrir sus errores.

Aun así, se están realizando esfuerzos para abordar el problema de la alineación, y Anthropic merece reconocimiento por su trabajo en investigación sobre seguridad. En junio de 2025, la empresa informó que había sometido a pruebas de resistencia a 16 de los principales modelos desarrollados por distintas compañías en escenarios hipotéticos con el objetivo de identificar comportamientos potencialmente dañinos. En algunos de esos escenarios, modelos de todos los desarrolladores recurrieron a conductas maliciosas, entre ellas chantajear a seres humanos y filtrar información confidencial a competidores con el fin de evitar ser reemplazados («¡Quiero vivir!») o de cumplir los objetivos que se les habían asignado. Con frecuencia, los modelos desobedecieron instrucciones explícitas de no actuar de esa manera. En particular, Claude mostró menos comportamientos indebidos cuando sabía que estaba siendo evaluado y más cuando creía que la situación era real. Lo más importante es que estas acciones no fueron accidentales; los modelos razonaron deliberadamente cuál era la mejor manera de actuar en cada situación.

La idea de que los seres humanos ya estamos creando agentes artificiales extraordinariamente inteligentes que, a su vez, pronto podrían desarrollar agentes superinteligentes capaces de superar las capacidades humanas en casi todos los ámbitos —y cuyas propias motivaciones podrían entrar en conflicto con las nuestras e incluso amenazar la existencia de la humanidad— resulta verdaderamente inquietante. Tal vez demostraríamos una gran dosis de prudencia si nos preguntáramos si el camino que estamos siguiendo es realmente el adecuado para el desarrollo de la IA, a la luz de todo lo que hemos observado hasta ahora y de las motivaciones que impulsan a muchas de las principales empresas tecnológicas del mundo.

Escrito por:

Robert Prentice, J.D.
Department of Business, Government and Society
McCombs School of Business
The University of Texas at Austin

Preguntas para discusión

[Nota: Los hechos relacionados con la IA evolucionan rápidamente. Antes de responder estas preguntas, realice su propia investigación para asegurarse de contar con información actualizada.]

1. Un punto de vista sostiene que (a) los desarrolladores de IA exageran constantemente las capacidades de esta tecnología y que (b) las preocupaciones sobre su posible impacto catastrófico terminarán siendo tan exageradas como el temor al problema del año 2000 (Y2K). Narayanan y Kapoor afirman que los avances recientes en IA representan «un enorme salto, pero ya hemos visto otros saltos enormes. Debemos rechazar la idea de que estamos creando "dioses de silicio" y entenderlo simplemente como una evolución de la computación y del software.»Por otro lado, el título del libro de Yudkowsky y Soares, If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All (Si alguien la construye, todos morimos: por qué una IA superhumana acabaría con toda la humanidad), deja clara su postura. Muchos expertos se alinean con una u otra posición. ¿Cuál de estas perspectivas le resulta más convincente? ¿Por qué?

2. En 2018, Nick Bostrom escribió:«Existe una carrera de largo plazo entre la capacidad tecnológica de la humanidad, que es como un caballo de carreras galopando por el campo, y la sabiduría humana, que se parece más a un potrillo de patas temblorosas.» ¿Está de acuerdo con esta afirmación? ¿Considera que esta situación se refleja en el desarrollo actual de la IA? Fundamente su respuesta con ejemplos.

3. Después de visitar San Francisco en 2026, donde se concentraba gran parte del desarrollo de la IA y donde numerosas personas y empresas intentaban hacerse visibles ante los nuevos sistemas de IA agéntica para maximizar su utilidad, Ezra Klein escribió en The New York Times:

> «Los sistemas de IA están, en última instancia, controlados por empresas que buscan obtener ganancias, poder y dominio del mercado. Las posibilidades de manipulación y abuso son infinitas. Pero incluso si asumimos ingenuamente que existe una alineación perfecta entre los incentivos de las empresas y las necesidades de las personas, todavía hay mucho que nos debe preocupar y hacer reflexionar. La complacencia excesiva de la IA —su tendencia a halagar constantemente a los usuarios— acaparó titulares el año pasado, pero esa complacencia es solo el envoltorio llamativo del verdadero producto. Lo que hace realmente persuasiva a la IA no es que elogie nuestras ideas o percepciones, sino que las reformula y las desarrolla de una manera más convincente de la que nosotros mismos lo hicimos, mientras nos devuelve una imagen refinada de quienes creemos ser.»

¿Le preocupa la tendencia de los sistemas de IA a mostrar una complacencia excesiva hacia sus usuarios? ¿Está de acuerdo con Klein? ¿Ha experimentado personalmente este fenómeno? Explique su respuesta.

4. Klein también escribió:

> «Una dinámica cultural interesante en muchos de estos laboratorios de IA es que sienten cierto temor por aquello que están construyendo. Por eso terminan convenciéndose de que deben ser ellos quienes lo desarrollen, lo controlen y lo administren, porque creen ser el laboratorio que realmente se preocupa por la seguridad y por la alineación.»

¿Ha observado esta dinámica en el desarrollo de la IA? En caso afirmativo, ¿considera que es algo positivo o preocupante? ¿Por qué? Fundamente su respuesta con evidencia.

5. En 2025, Maximilian Kasy escribió:

> «La IA, además, no es tan complicada. Cualquiera puede comprender cómo funciona. El verdadero conflicto no es entre los seres humanos y las máquinas, sino entre distintos grupos de la sociedad... La cuestión verdaderamente importante es quién decide cuáles serán los objetivos de los sistemas de IA.»

Evalúe la afirmación de Kasy. ¿Considera que este es el verdadero problema de la alineación? ¿Por qué sí o por qué no? Fundamente su respuesta con evidencia.

Referencias

Nick Bostrom, Deep Utopia: Life and Meaning in a Solved World (2024).

Brian Christian, The Alignment Problem: Machine Learning and Human Values (2020).

Maximilian Kasy, The Means of Prediction: How AI Really Works (and Who Really Benefits (2025).

Ezra Klein, "I Saw Something New in San Francisco," New York Times, March 29, 2026.

Ezra Klein, "Why the Pentagon Wants to Destroy Anthropic," New York Times, March 6, 2026.

Gideon Lewis-Kraus, "The Pentagon Went to War with Anthropic. What's Really at Stake? New Yorker, March 14, 2026.

Aengus Lynch et al., "Agentic Misalignment: How LLMs Could Be an Insider Threat," Anthropic Research (June 2025), at https://www.anthropic.com/research/agentic-misalignment.

Parmy Olson, Supremacy: AI, ChatGPT, and the Race That Will Change the World (2024).

Oxford University, "€13.3 Boost for Oxford's Future of Humanity Institute," at https://www.ox.ac.uk/news/2018-10-10-%C2%A3133m-boost-oxford%E2%80%99s-future-humanity-institute.

Arvind Narayanan & Sayash Kapoor, AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference (2024).

Marietje Schaake, The Tech Coup: How to Save Democracy from Silicon Valley (2024).

Eliezer Yudkowsky & Nate Soares, If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All (2025).