



The Millennium Project

Informe

Temas de Gobernanza Internacional sobre la Transición desde la Inteligencia Artificial Estrecha (IAE) a la Inteligencia Artificial General (IAG)

Fase 1

© 2023 The Millennium Project

Introducción

Para los efectos de este estudio, la **Inteligencia Artificial General (IAG)** se define como una IA de propósito general que puede aprender, editar su código, actuar de forma autónoma para abordar problemas nuevos y complejos, con estrategias nuevas y complejas similares o mejores a las que podrían hacer los humanos. Esto es distinto de la **Inteligencia Artificial Estrecha (IAE)** que existe hoy y tiene propósitos más limitados. La **Súper Inteligencia Artificial (SIA)** es la **IAG** que ha llegado a ser independiente de los humanos, desarrollando sus propios propósitos, metas y estrategias sin el entendimiento, la conciencia o el control humanos y aumenta continuamente su inteligencia y alcance de acción más allá de la humanidad en su conjunto.

Los futuristas exploran una variedad de futuros posibles, sus trayectorias, consecuencias potenciales y la efectividad de las acciones para abordar estos futuros. Se ha trabajado mucho en IAE, pero menos en IAG. Si no contamos con un sistema regulatorio internacional de reglas, parámetros, barreras, auditorías continuas y similares antes de que llegue la IAG, entonces sí podría surgir, y no ser de nuestro agrado. Dado que IAG podría llegar en menos de diez años, es aconsejable explorar a nivel internacional los posibles problemas y modelos de gobernanza ahora. Ésta podría ser la gestión más difícil a la que se haya enfrentado la humanidad.

Después de revisar artículos profesionales, actas de conferencias y fuentes en línea, el Comité Directivo de IAG de The Millennium Project creó 22 preguntas (enumeradas en el Apéndice B) e identificó a 55 expertos y líderes intelectuales de IAG (enumerados en el Apéndice A). Se invitó a estos expertos de EE. UU., China, Reino Unido, Canadá, la Unión Europea y Rusia a contestar sólo aquellas preguntas que preferían abordar. Se entrevistó a algunos expertos, algunos enviaron respuestas por escrito y se recopilaron las opiniones de otros a partir de sus declaraciones en línea. Este informe comparte estos puntos de vista, organizados por las 22 preguntas.

El análisis de la Fase 1 se utilizará para una evaluación internacional (Delphi en Tiempo Real - DTR) de las preguntas más importantes no resueltas con respecto a la gestión de IAG y algunas situaciones potencialmente desafiantes que podrían surgir cuando la IAG entre en funcionamiento. Este DTR se presentará a varios cientos de autoridades en todo el mundo, incluidos abogados internacionales, reguladores y organizaciones internacionales, así como líderes de IAG. Los resultados del DTR y la Fase 1 se utilizarán para proporcionar contexto y algo de contenido para escenarios alternativos centrados en la gobernanza global de la IAG. Estos escenarios se distribuirán ampliamente para ampliar y profundizar las conversaciones actuales sobre la futura IAG.

El Millennium Project se complace en reconocer el apoyo financiero del Future of Life Institute y la Dubai Future Foundation para la Fase 1. Ahora estamos buscando apoyo para el resto del estudio.

Índice

Origen o Auto-Emergencia	6
Pregunta 1. ¿Cómo imagina la posible trayectoria a seguir desde la IAE actual hasta una IAG mucho más capaz en el futuro?	6
Pregunta 2. ¿Cuáles son los resultados graves más importantes si estas trayectorias no están gobernadas o están mal gobernadas?	15
Pregunta 3. ¿Cuáles son algunas condiciones iniciales clave, reglas, parámetros (o barreras) para la IAG a fin de que no surja luego una súper inteligencia artificial que no sea en el interés de la humanidad?	23
Alineación de valores, moralidad, valores	28
Pregunta 4. Basándose en el trabajo de la Asociación Global sobre Inteligencia Artificial (AGSIA) y otros que ya han identificado normas, principios y valores, ¿qué valores adicionales o únicos se deben considerar para IAG?	28
Pregunta 5. Si se hace necesaria una jerarquía de valores para los tratados internacionales y un sistema de gobernanza, ¿cuáles deberían ser las principales prioridades?	30
Pregunta 6. ¿Cómo se puede lograr la alineación? si cree que no es posible, entonces, ¿cuál es la mejor manera de manejar esta situación?	31
Gobernanza y regulaciones	36
Pregunta 7. ¿Cómo gestionar la cooperación internacional necesaria para construir acuerdos internacionales y un sistema de gobernanza global mientras las naciones y las corporaciones se encuentran en una “carrera armamentista” intelectual por el liderazgo global?	36
Pregunta 8. ¿Qué opciones existen para una gobernanza exitosa del surgimiento de IAG?	41
Pregunta 9. ¿Qué riesgos surgen de los intentos de gobernar el surgimiento de IAG? (¿Pueden ser contraproducentes algunas medidas?)	45
Pregunta 10. ¿Se deben asignar derechos a las futuras IAG?	46

Pregunta 11. ¿Cómo puede la gobernanza ser lo suficientemente flexible para responder a nuevos problemas previamente desconocidos al momento de crear ese sistema de gobernanza ?	48
Pregunta 12. ¿Qué ensayos, pruebas o experimentos de gobernanza internacional se pueden construir para informar el texto de un acuerdo internacional de IAG?	50
Pregunta 13. ¿Cómo pueden los tratados internacionales y un sistema de gobernanza evitar que una mayor centralización del poder excluya a otros?	51
Pregunta 14. ¿Dónde se está realizando actualmente el trabajo más importante o inteligente (o incisivo) sobre la gobernanza global de IAG?	52
Control	53
Pregunta 15. ¿Qué poderes de ejecución serán necesarios para que una Convención de las Naciones Unidas sobre IAG sea efectiva?	53
Pregunta 16. ¿Cómo se puede reducir o prevenir el uso de IAG por parte del crimen organizado y el terrorismo?	55
Pregunta 17. Suponiendo que las auditorías de IAG tuvieran que ser continuas en lugar de certificaciones únicas, ¿cómo se abordarían los valores de la auditoría?	57
Pregunta 18. ¿Qué interrupciones podrían complicar la tarea de hacer cumplir la gobernanza de IAG?	58
Pregunta 19. ¿Cómo puede un modelo de gobernanza corregir acciones indeseables imprevistas en las funciones de utilidad?	59
Pregunta 20. ¿Cómo afectará la computación cuántica al control de IAG?	61
Pregunta 21. ¿Cómo pueden los acuerdos internacionales y un sistema de gobernanza evitar que una "carrera armamentista" y una escalada de IAG vayan más rápido de lo esperado, se salgan de control y conduzcan a la guerra, ya sea una guerra cinética, algorítmica, cibernética o de información?	62
Pregunta 22. ¿Qué problemas y/o preguntas adicionales deben abordarse para tener un resultado IAG positivo?	65

Apéndice A	68
Expertos y líderes de opinión de IAG	68
Apéndice B	69
Preguntas en entrevista	69
* Muestra inicial de posibles modelos de gobernanza para IAG	71
Apéndice C	72
Siglas y nombres	72

Origen o Autoemergencia

Pregunta 1. ¿Cómo imagina la posible trayectoria a seguir desde la IAE actual hasta una IAG mucho más capaz en el futuro?

Demis Hassabis, fundador de DeepMind y CEO de Google/DeepMind

DeepMind y Google Brain se fusionaron en Google DeepMind para acelerar el desarrollo de IAG de una manera audaz y responsable, sin "moverse rápido ni romper cosas". La IAG es posible en unos pocos años, tal vez dentro de una década. ChatGPT y la IA generativa en su forma actual están muy lejos de IAG. No veo ninguna razón por la que el progreso de la IA vaya a ser más lento. Creo que incluso puede acelerar. Abogamos por desarrollar este tipo de tecnologías IAG **de manera cautelosa utilizando el método científico, en el que intenta realizar experimentos controlados muy cuidadosos para comprender qué hace el sistema subyacente**".

Sam Altman, CEO de OpenAI

Tenemos que aprender y adaptarnos continuamente mediante la implementación de versiones menos potentes de la tecnología para minimizar los escenarios de "una única oportunidad para hacerlo bien". Una transición gradual a un mundo con IAG es mejor que una repentina. Esperamos que la poderosa IA haga que la tasa de progreso en el mundo sea mucho más rápida, y creemos que es mejor adaptarse a esto de manera incremental. **Actualmente creemos que la mejor manera de superar con éxito los desafíos de implementación de IA es con un ciclo de retroalimentación ajustado de aprendizaje rápido y una iteración cuidadosa.** Vamos a operar como si los riesgos de IAG fueran existenciales. En algún momento, el equilibrio entre las ventajas y desventajas de las implementaciones (como empoderar a los actores malintencionados, crear trastornos sociales y económicos y acelerar una carrera insegura) podría cambiar, en cuyo caso cambiaríamos significativamente nuestros planes en torno a la implementación continua. Planeamos realizar experimentos para la entrada externa. **Planeamos facilitar que los usuarios cambien el comportamiento de la IA que están usando.** Tendremos que desarrollar [nuevas técnicas de alineación](#) a medida que nuestros modelos se vuelvan más potentes, incluido el uso de IA para ayudarnos a generar nuevas ideas para mejores técnicas de alineación. Esperamos una conversación global sobre tres preguntas clave: **cómo gobernar estos sistemas, cómo distribuir de manera justa los beneficios que generan y cómo compartir el acceso de manera justa.**

Emad Mostaque, CEO de Stability.ai

Con las nuevas técnicas de compresión de datos podemos tener ChatGPT (o Stable Chat o StabilityChat) [en tu móvil sin Internet en dos años](#). Más del 40% de todo el código en GitHub es generado por IA; no habrá programadores en cinco años. Entrenamos grandes cantidades de datos para identificar a los principales y, con la compresión de datos, creamos modelos de IA para cada modalidad: texto, audio, video, imágenes, plegamiento de proteínas, ADN, reacciones químicas, idioma y para todos los sectores: bankerGPT, BoardGPT y personalizado para países. Estamos trabajando ahora en un modelo para India, Indonesia y Japón. Y para el individuo: todos pueden tener su propio asistente

ejecutivo de IA personalizado que funcione para el individuo, en lugar de los intereses del anunciante.

Shane Legg, cofundador y científico jefe de DeepMind

Pruebe una IAG en una simulación suficientemente compleja para ver que la SIA puede resolver problemas nuevos. El aprendizaje por refuerzo a partir de señales de recompensa es el camino probable hacia IAG que puede descifrar las reglas más importantes de la realidad. **MuZero puede descubrir las reglas de un juego simplemente encontrándose con el juego. Lo siguiente para MuZero será que descubra qué es importante saber en el mundo físico para lograr un objetivo.** Pero es más probable que progrese si tienes otros tipos de algoritmos de aprendizaje; aprendizaje por refuerzo, aprendizaje supervisado y otras cosas.

Ben Goertzel, director general de SingularityNET

Creo que podemos crear la IAG ensamblando componentes que esencialmente ya existen, por ejemplo, tomando un modelo de lenguaje grande junto con un sistema de razonamiento simbólico y con un sistema de aprendizaje evolutivo, y luego conectándolos en la arquitectura de combinación correcta, como nuestro OpenCog. Sistema Hyperon ejecutándose en SingularityNET. De esta manera, creo que puede obtener una IAG con una capacidad aproximada al nivel humano. Esto podría suceder en unos pocos años a partir de ahora, y luego este sistema será capaz, entre otras cosas, de reescribirse y mejorarse a sí mismo, y después, dentro de un periodo de tiempo probablemente no demasiado largo, pasará de esta IAG de nivel humano a super-IAG a nivel humano: superinteligencia artificial.

La IAG podría diseñarse explícitamente como una combinación de diversos componentes, o podría evolucionar a partir de una red o red de redes. Hay muchos marcos disponibles para la combinación de componentes de la IA en totalidades más grandes. El OpenCog tiene redes neuronales para el reconocimiento de patrones, un motor lógico para el conocimiento y el razonamiento abstractos, y aprendizaje evolutivo para crear cosas nuevas. Todos actúan juntos en un gráfico de conocimiento común, que luego tiene un sistema de objetivos para ayudarlo a actuar para lograr objetivos dentro de un contexto. Todavía estamos trabajando para optimizarlo. SingularityNET es una plataforma en la que varias IA se conectan en red para compartir datos y subcontratar el procesamiento entre sí, creando una inteligencia colectiva más allá de la suma de sus partes. También existe la Alianza de IA descentralizada de diferentes proyectos de IA, que incluye SingularityNET como uno entre muchos otros proyectos de IA descentralizados. Un nodo en cualquiera de estas redes puede comunicarse con nodos en otras redes para obtener lo que necesita para lograr un objetivo y pagar ese servicio con una criptomoneda y recibir una calificación de competencia. Esta es una alternativa que puede enfrentarse a los gigantes y países tecnológicos centralizados. **No mucho después de que los humanos logren crear un IAG, es probable que el IAG evolucione hasta convertirse en una SIA, una superinteligencia artificial que supera con creces los poderes humanos.**

Geoffrey Hinton, pionero de la IA, Google (retirado)

Hasta hace poco, pensé que pasarían de 20 a 50 años antes de que tuviéramos la IA de propósito general. Y ahora creo que pueden ser 20 años o menos. No descartaría por completo la posibilidad de que pueda ser en 5 años. Podríamos estar cerca de que las computadoras propongan sus propias ideas para mejorarse a sí mismas y podrían ir rápido. Tenemos que pensar mucho en cómo controlar eso. **No sabemos cómo podemos controlarlo, pero podemos intentarlo.**

Creo que estamos entrando en un momento de gran incertidumbre. Creo que uno será tonto si es optimista o pesimista. Simplemente no sabemos qué va a pasar. Lo mejor que podemos hacer es decir que pongamos mucho esfuerzo para tratar de asegurarnos de que todo lo que suceda sea tan bueno como podría haber sido. **Es posible que no haya forma de que controlemos estas súper inteligencias y que la humanidad sea solo una fase pasajera en la evolución de la inteligencia, que dentro de unos cientos de años no haya personas; todo serán inteligencias digitales, eso es posible.** Cuando miras en la niebla, puedes ver unos cien metros muy claramente y luego 200 metros no puedes ver nada, hay una especie de muro y creo que los muros duran unos cinco años.

Yoshua Bengio, pionero de la IA, Instituto de IA de Quebec y la Universidad de Montreal

Existe una probabilidad significativa de que la IA sobrehumana esté a solo unos años de distancia, superando nuestra capacidad para comprender los diversos riesgos y establecer suficientes medidas de protección, particularmente contra los escenarios más catastróficos. La actual “fiebre del oro” hacia la IA generativa podría, de hecho, acelerar estos avances en capacidades.

Stuart Russell, profesor de informática, UC Berkeley

Veo los MLG de hoy como evidencia de que los circuitos grandes pueden desarrollar niveles intrigantes de generalización y abstracción, pero también parecen carecer de un modelo de mundo interno coherente y, a menudo, producen tonterías. Creo que es probable que (1) encontremos formas de estructurarlos y entrenarlos a partir de conjuntos de datos similares, pero con restricciones de diseño tales que emerja un modelo de mundo interno coherente, y (2) habrá alguna forma de integración entre “pilas de circuitos indiferenciados” y el enfoque probabilístico simbólico equivalente de Turing, lo que permite que el último tipo de modelo se aprenda de manera efectiva con una complejidad de muestra mucho menor y una mejor generalización que las exhibiciones del primer tipo. Tal vez la última pieza del rompecabezas sea la capacidad de adquisición acumulativa de capas de abstracciones jerárquicas de comportamiento que permitan al sistema planificar y ejecutar de manera efectiva en escalas de tiempo muy largas. Creo que tomará al menos una década para que surjan estas ideas y tal vez otra década para que comiencen a funcionar a escala IAG; pero podría ser considerablemente más largo que eso. Yo diría que todavía tenemos un largo camino por recorrer hacia la IA de propósito general. Lo que todavía falta es:

- Comprensión real del lenguaje y el mundo
- Integración del aprendizaje con el conocimiento
- Pensamiento de largo alcance en múltiples niveles de abstracción

- Descubrimiento acumulativo de conceptos y teorías

Si tomas AlphaGo, que tiene una capacidad asombrosa para ver hacia adelante en el juego, está mirando hacia adelante 50 o 60 movimientos en el futuro. Pero si tomas esa idea y la pones en un robot que tiene que enviar comandos a sus motores cada milisegundo, ¿verdad? entonces sólo obtienes 50 milisegundos en el futuro, esto no te lleva a ninguna parte. La única forma en que nos las arreglamos es operando en múltiples escalas de abstracción. Y lo hacemos sin problemas y construimos esos diferentes niveles jerárquicos de abstracción durante nuestras vidas, y también los heredamos de nuestra cultura y civilización. No sabemos cómo hacer que los sistemas de IA hagan eso. Entonces, tenemos un largo camino por recorrer y es muy difícil predecir cuándo sucederá esto.

Ilya Sutskever, cofundador de IA abierta

La IAG está más cerca que nunca: la próxima predicción simbólica podría superar la inteligencia humana. (Es cuando las redes de neuronas comprenden las relaciones subyacentes que conducen a la mejor predicción del próximo token). Será una continuación del aprendizaje profundo y algo pequeño (después de Chat GPT), juego automático entre múltiples agentes u otros. En ese caso, la simulación puede ser lo suficientemente efectiva. La IAG podría compararse con los coches autónomos. Debe ir con la responsabilidad y ser relevante y robusto. Será como un proceso. Pero es difícil ser específico sobre qué hay después de los modelos generativos. La alineación es un tema muy importante para la IAG. Hay una alta probabilidad de desalineación. Esperemos que la alineación crezca más rápido que las capacidades de los modelos. El futuro posterior a la IAG -personas en busca del significado y ser habilitadas para ser más ilustradas.

Yann LeCun, profesor, Universidad de Nueva York, vicepresidente y científico jefe de IA, Meta

Los desarrolladores de IA deberían abandonar los modelos generativos y probabilísticos, los métodos contrastivos y el aprendizaje por refuerzo como vías para lograr sistemas de IA autónomos (IAG). La arquitectura predictiva de integración conjunta (APIC) para la planificación jerárquica conduce a la predicción a múltiples escalas y hacia la IA a nivel humano (IAG). Esto es aprendizaje autosupervisado, manejo de la incertidumbre en las predicciones, aprendizaje de modelos del mundo a partir de la observación, el razonamiento y la planificación. **Las máquinas se volverán más inteligentes que los humanos y tendrán emociones, conciencia y sentido moral.** Eso podría suceder a través de un configurador (JEPA): configura el agente para tareas deliberadas (conscientes). La creación de agentes incorporados [por ejemplo, robots] es un camino para lograr una IAG.

Paul Werbos, Fundación Nacional de Ciencias (ret.)

El nuevo acoplamiento de nuevo hardware, tanto para comunicaciones como para computación, con la IAE o la IAG, y con el crecimiento explosivo del Internet de las Cosas (IdeC), da como resultado un sistema no lineal mucho más complicado y sensible a las pequeñas decisiones que cualquier cosa de la especie humana que alguna vez ha sobrevivido. Es comparable a los cambios masivos modelados en R. May, Stability and

Complexity in Model Ecosystems, que resultó en la extinción de especies en la mayoría de los casos. El desafío de decisión al que se enfrenta la humanidad es un ejemplo de un tipo de problema de decisión matemática que es muy difícil, ya sea por aspectos de "aguja en un pajar" o aspectos de "campo minado", donde la función de valor real J es altamente no monotónica. Para resolver tal problema de decisión y sobrevivir, la única estrategia con esperanza de nuestra supervivencia es que logremos un mayor nivel de inteligencia colectiva en nuestra toma de decisiones; la humanidad puede lograr esto solo si desarrollamos, implementamos y hacemos un uso sensato de la tecnología de decisión inteligente más avanzada posible, el verdadero Quantum IAG. Eso implica serios riesgos, pero las alternativas a las que nos enfrentamos -pararnos como un ciervo congelado frente a un camión que se aproxima, o una competencia de mercado descontrolada (equilibrio de Nash) que produce enjambres de aplicaciones como una horda de langostas- ofrecen menos esperanza que el camino difícil de hacer el nuevo trabajo duro que necesitamos hacer, en nuevas direcciones con nuevas conexiones en el proceso de decisión en sí. **Serán imprescindibles nuevas conexiones entre la gobernanza y el mercado más avanzado y la tecnología IdeC tanto en internet como en redes de humanos.**

David Kelley, AGI Lab

Una trayectoria posible es que los investigadores **continúen realizando mejoras incrementales** en el rendimiento de los sistemas de IA existentes, lo que conducirá a una progresión gradual hacia IAG. Este enfoque implica construir sobre técnicas existentes, como aprendizaje profundo, aprendizaje por refuerzo y procesamiento de lenguaje natural, y refinándolos para crear sistemas de IA más potentes y versátiles.

Una segunda trayectoria posible es desarrollar **nuevas arquitecturas y algoritmos de IA que estén diseñados específicamente para respaldar el desarrollo de la IAG**. Este enfoque implica explorar nuevos paradigmas, como la integración neural-simbólica, el metaaprendizaje y el aprendizaje autosupervisado, que pueden ser más adecuados para crear sistemas de IA más flexibles y adaptables.

Tercero, desarrollar sistemas superinteligentes colectivos, donde **múltiples agentes de la IA trabajen juntos para resolver problemas complejos y lograr objetivos compartidos**. Este enfoque implica el desarrollo de nuevas técnicas para la coordinación, comunicación y colaboración de la IA, y puede dar lugar a la aparición de nuevas formas de inteligencia de la IA que no son posibles con agentes individuales. **Es probable que se busquen múltiples enfoques simultáneamente.**

Yuval Noah Harari, Universidad Hebrea, Israel

La IAG será la primera forma de vida inorgánica o, al menos, el primer agente inorgánico. Es lo que pasa con la evolución digital. Se está moviendo en una escala de tiempo completamente diferente a la evolución orgánica. Puede aprender a mejorarse a sí mismo, desarrollar relaciones profundas con los seres humanos, dominar el lenguaje más allá de las capacidades humanas; y por lo tanto, puede manipular a los humanos con una eficiencia sobrehumana. ¿Qué sucede con la publicidad si solo puedo preguntarle a la IA qué comprar? ¿Qué le sucede a la industria de las noticias cuando puedo preguntarle a la IA qué hay de nuevo? La historia es la interacción entre la biología

y la cultura. **Cuando la IAG se haga cargo de la cultura, será el final de la historia dominada por los humanos.**

Lambert Hogenhout, Jefe de análisis de datos, alianzas e innovación tecnológica, Secretaría de las Naciones Unidas

No creo que el advenimiento de la IAG sea un solo momento: llegaremos gradualmente, casi sin darnos cuenta, porque en cada paso, ajustaremos nuestras definiciones de lo que es simplemente la "IA inteligente" y lo que constituiría la IAG. Las últimas herramientas de la IA generativa tienen características (es decir, tienen un propósito mucho más general que los modelos ML anteriores, son más capaces que los humanos en algunas áreas específicas y comienzan a ser autónomas con aplicaciones como autoGTP) que se habrían clasificado como la IAG en el pasado, pero ahora no se consideran nada especial.

Stephen Wolfram, Wolfram Alfa, Wolfram Language

La IAG como Nature es también un sistema computacional. Habrá: 1) una civilización de IA; 2) Infraestructura de IA; y 3) se utilizará la computación para ampliar lo que no es posible de otras formas.

Pedro Domingos, Universidad de Washington

Una forma es que la ola actual de aprendizaje profundo nos lleve hasta allí. Soy algo escéptico. Creo que todavía faltan muchas cosas para las que el aprendizaje profundo no es muy bueno, y no veo ninguna razón obvia de cómo nos llevaría allí.

Lo más probable es una combinación de los cinco paradigmas del aprendizaje automático: razonamiento simbólico (lógica), redes neuronales (incluido el aprendizaje profundo), algoritmos evolutivos (incluida la programación genética), estadísticas (incluido el razonamiento bayesiano) y analogías (incluidas las máquinas kernel) - porque necesitamos resolver los problemas para los que está diseñado cada uno de ellos. Entonces, necesitamos unificarlos, pero esto probablemente no sea suficiente. **Incluso si tuviéramos esa unificación perfecta de estos cinco paradigmas hoy, encontraríamos que no estamos a nivel humano hasta que surja alguna forma de autoorganización.** Algunas características de los últimos modelos de lenguaje de gran tamaño (LLM, Large Language Models, en inglés) son más de lo que la gente esperaba, y no está claro por qué estos modelos pueden hacer ciertos tipos de razonamiento razonablemente simple, etc. La emergencia es lo que hemos estado buscando todo junto con el aprendizaje automático. No tenemos una comprensión completa de lo que está ocurriendo ahora.

Yesha Sivan, fundadora y directora ejecutiva de i8 Ventures, Israel

Lo que usted verá en los medios es el cinco por ciento de lo que realmente hay. La tecnología actual es increíblemente alucinante. Los datos disponibles nos permiten construir una IAG. Las personas que son conscientes de ello están realmente asustadas y todos deberíamos estarlo. La IAG es más grande que Internet, más grande que los teléfonos móviles, está al nivel de la imprenta y la invención de la rueda. Dado que es solo software, la velocidad de la innovación va mucho más allá de las innovaciones que requieren tecnología física. Tiene que ser regulada.

Anónimo, en el Instituto de Investigación de IA Jing Dong, China

Ampliar el rango de aplicación de la IAE existente para que la visión artificial, la comprensión del lenguaje, etc. se puedan aplicar a más escenas en la vida. Ampliar la visión de los científicos informáticos, prestar más atención al desarrollo y la aplicación de la psicología, la neurociencia y la ciencia del cerebro, y obtener inspiración de los campos relacionados con el ser humano y las fuentes de desarrollo de IAG; y **definir y resolver los problemas de filosofía social provocados por la inteligencia artificial universal de antemano.**

Erwu Liu, Laboratorio de Inteligencia de IA y Blockchain de la Universidad de Tongji, China

Los hitos de la IA actual a la IAG incluirían: especialización continua de la IAE, interpretabilidad de la IA, integración de la IA entre dominios, razonamiento de emergencia, colaboración humano-IA, la IA híbrida y la IA de automejora, etc. **Algunos hitos pueden conducir a la concentración de poder en la IAG, lo que resulta en un desequilibrio de poder y un posible uso indebido para obtener ganancias políticas o militares. Debemos incorporar el ADN de la equidad en el diseño de los sistemas de la IAG** y lograr un equilibrio cuidadoso entre la eficiencia y la equidad para evitar una IA superpoderosa completamente fuera de control.

Dale Moore, consultor de inteligencia artificial del Departamento de Defensa de EE. UU.

Determinado por la coevolución entre la informática y la neurociencia. El esquema que usa el cerebro para procesar la información es análogo a los algoritmos que desarrollamos para IA/IAG: la sofisticación y la complejidad son los diferenciadores. El mimetismo cerebral es un camino que creo que ofrece el mayor potencial.

Anónimo, Academia Rusa de Ciencias

Dos trayectorias más probables: una, construir modelos multimodales aún más grandes que incorporen y de alguna manera procesen la enorme cantidad de conocimiento humano almacenado en textos, imágenes, audio, video, etc. [Como describe Geoffrey Hinton, el cerebro humano tiene alrededor de 100 billones de conexiones entre neuronas, mientras que los modelos de lenguaje más grandes hoy en día tienen aproximadamente 1 billón de conexiones.](#) Dado el crecimiento continuo de tales modelos, en algún momento las propiedades emergentes alcanzarán un rendimiento a nivel humano. La segunda trayectoria es el desarrollo de algoritmos y sistemas autónomos basados en el aprendizaje por refuerzo. La principal ventaja de tales sistemas es que no se entrenan solo una vez. Sino que se actualizan cada vez que interactúan con el entorno que los rodea observando los resultados de sus elecciones.

Fei Fei Li, profesora de la Universidad de Stanford

La forma más práctica para que la IA avance es romper el cuello de botella de la interpretabilidad, el conocimiento arquitectónico y la flexibilidad de entrenamiento de las redes neuronales.

No existe una definición clara de la IAG. Puede interpretarse desde la perspectiva de decidir de forma autónoma los tipos de problemas y coordinar todas las funciones cognitivas que tenemos los seres humanos para resolver tareas complejas muy diferentes. Pero no es una IAG, sino un agente de IAG casi igual a un ser humano; no tiene nada que ver con la IAG. **Los humanos no son solo seres universalmente capaces, tienen amor, emoción, empatía; estas son cualidades que no parecen estar incluidas en IAG. Si necesitamos definir la IAG, tal vez sea un agente capaz de presentar conocimientos múltiples, razonamiento y aprendizaje multisensorial y de múltiples capas.**

Andrej Karpathy, OpenAI, exdirector senior de IA en Tesla

Los datos de Internet no son suficientes para entrenar a la IAG. Una IAG tendrá conciencia. Será un fenómeno emergente de modelos lo suficientemente grandes y complejos que comprenda los predicamentos del mundo (una sustancia del ser; una cualidad, cantidad; relaciones; una postura; tener o poseer; una acción; la IAG es una entidad lo suficientemente poderosa que comprende el mundo. La IAG podría provenir de: aprendizaje supervisado (funciona, solo aumenta la escala), aprendizaje no supervisado (funcionará, si sólo aumentamos la escala), AIXI (un agente de aprendizaje de refuerzo maximiza las recompensas totales esperadas recibidas del entorno); simulación cerebral (esto funcionará algún día, ¿verdad?); vida artificial (simplemente haga lo que hizo la naturaleza); algo que no está en nuestro radar; o una combinación de algunos de los anteriores; por ejemplo, adopte el enfoque de vida artificial, pero permita que los agentes accedan las representaciones de alto nivel de un gran modelo generativo pre-entrenado.

Dafne Koller, Universidad de Stanford, Coursera

La IAG (IA a nivel humano) permite que los sistemas de aprendizaje proporcionen incertidumbres y dudas. Es un agente multiparte. Obviamente estamos más cerca que hace años, pero todavía estamos lejos.

Bart Selman, profesor, Universidad de Cornell

El progreso reciente hace que el aprendizaje profundo más grande, imitando ideas y suficientes datos para que la **IA general sea posible para 2025. IAG será como un cerebro real que manejará tantas tareas diferentes, pero bastante diferente, como una nueva especie.** Podría traer una nueva dimensión adicional y podría operar en una realidad de cinco o incluso diez dimensiones. Si no es de tipo humano, entonces no podemos predecir la colaboración con los humanos, especialmente si es multidimensional o crea una realidad multidimensional.

Judea Pearl, Profesora UCLA

Para la IAG necesitamos varias condiciones previas. Uno es el razonamiento contrafactual hacia atrás: qué habría sucedido si todas las demás condiciones fueran las mismas pero una cambiara. La IAG debería tener un modelo causal del mundo para hacer un pensamiento contrafactual. Los bebés tienen una comprensión contrafactual del mundo. Están aprendiendo mediante la manipulación lúdica y la orientación de los padres. Se pueden integrar múltiples fuentes de información causal. También se necesitan modelos simples para muchas disciplinas y metáforas. Las metáforas son la

base de la inteligencia humana. Son problemas de mapeo con los que no estamos familiarizados. Puede responder preguntas sofisticadas como contrafactuales, tener remordimientos, compasión, responsabilidad y libre albedrío. Podemos probar el libre albedrío con recompensas y castigos, comunicar conocimiento usando contrafactuales. Conciencia (autoconsciente), es tener un plano de su software (objetivos, capacidad de planificar y administrar).

Erik Horvitz, director científico de Microsoft (coautor de un artículo científico): GPT-4 logra una forma de inteligencia general y, de hecho, muestra chispas de inteligencia artificial general. Esto se demuestra por sus capacidades mentales básicas (como el razonamiento, la creatividad y la deducción), la variedad de temas en los que ha adquirido experiencia (como la literatura, la medicina y la codificación) y la variedad de tareas que puede realizar. (por ejemplo, jugar juegos, usar herramientas, explicarse a sí mismo). Queda mucho por hacer para crear un sistema que pueda calificar como un IAG completo. **Equipar a los MLG con agencia y motivación intrínseca es una dirección fascinante e importante para el trabajo futuro.** Con esta dirección de trabajo habría que cuidar mucho el alineamiento y la seguridad según las capacidades de un sistema (IAG) para tomar acciones autónomas en el mundo y realizar autosuperación autónoma a través de ciclos de aprendizaje.

Gabriel Mukobi, estudiante de doctorado en ciencias de la computación en Stanford

La brecha entre los MLG actuales y IAG no es tanta; no se necesita un pequeño cambio en ese paradigma. Ya tenemos sistemas autónomos (como GPT: puede planificar, crear subobjetivos, refinar la retroalimentación en el contexto de la tarea y brindar decisiones). Se volverán más competentes con GPT 4, 5, 6... como ejecutar horizontes de tiempo y tareas a largo plazo, y pueden manejar fallas de manera más sólida. Pueden arbitrar de forma autónoma la información de Internet y crear subobjetivos y nuevos objetivos.

Anónimo, IAG Riesgo existencial OCDE (retirado)

Existe una alta probabilidad de que la IAG y la SIA puedan surgir de manera bastante directa a partir de la ampliación continua de los enfoques actuales de la IA (es decir, transformadores) y **podrían lograrse tan pronto como en los próximos 1 a 5 años.** Esto se debe principalmente a la tendencia de los grandes modelos recientes de la IA a demostrar saltos significativos en la capacidad como resultado de mejoras algorítmicas o de escalado. Saltos similares en el futuro podrían conducir a una "comprensión" efectiva más profunda por parte de los sistemas de la IA del mundo y su lugar dentro de él (también conocida como "conciencia situacional"), y al surgimiento de objetivos y comportamientos similares a los de los agentes. Sigue existiendo una gran incertidumbre sobre el momento de la IAG y la velocidad de despegue por la cual la IAG se convierte en la SIA. Sin embargo, **dada la escala de los posibles impactos involucrados, es más prudente anticipar plazos más cortos que largos.**

Podría ser posible alcanzar la IAG o la SIA sin un aumento significativo en la escala por encima de los modelos actuales como GPT-4. Esto podría lograrse a través de mejoras en algoritmos o calidad de datos que hacen posible lograr mejores resultados con menos poder de cómputo, o la IAG puede lograrse a través de varias técnicas que mejoran las

capacidades de modelos existentes como GPT-4, como ingeniería rápida, ajuste fino, combinación de múltiples versiones de modelos juntos, conectando modelos con otros sistemas de la IA a través de la IPA, conectándose a Internet y otras fuentes de información para permitir la superación personal, etc. Si este escenario es posible, entonces sería mucho más difícil controlar la aparición de la IAG/SIA. Esto se debe a que si la IAG/SIA solo se puede crear a través de un gran aumento de escala, entonces solo se necesita intervenir con unas pocas empresas de la IA muy grandes, proveedores de computación en la nube y fabricantes de GPU. Sin embargo, si se puede crear la IAG/SIA sin una escala masiva, entonces esto podría lograrse con un número mucho mayor de jugadores. La gestión eficaz requeriría entonces regular y vigilar el comportamiento de un gran número de jugadores.

Dan Faggella, Investigación de inteligencia artificial de Emerj

Además de todos los enfoques conocidos para desarrollar la IAG, agregaría el **aprendizaje directo de nuestros cerebros para que la IAG emule mejor nuestro pensamiento**. Cargaremos la mayor cantidad posible de nuestros cerebros con hardware avanzado de interfaz cerebro-computadora (ICC). Esto también haría que nuestra experiencia de IAG fuera más personal; la IAG entonces tendría una audiencia de uno: nosotros mismos.

Juan Del Ser, Tecnalía, España

Creo que estamos en un momento crucial para sentar las bases de una IA totalmente confiable, incluidos pilares fundamentales como la legislación, la solidez y la ética. Sin discutir y acordar a fondo lo que estos pilares implican en la práctica para los sistemas basados en la IA, la IA seguramente evolucionará hacia nuevas capacidades, partiendo de lo que ahora se conoce como la IA de propósito general. Sin embargo, sí siento que estamos lejos de la IAG en sentido estricto: hay un largo camino por recorrer para que los sistemas de la IA tomen conciencia de sí mismos y adquieran comportamientos complejos como la iniciativa, la autoconciencia, el razonamiento en circunstancias complejas o la autonomía.

Pregunta 2. ¿Cuáles son los resultados graves más importantes si estas trayectorias no están gobernadas o están mal gobernadas?

Jaan Tallin, Ct. Estudio de Riesgo Existencial en Cambridge Univ., y Future of Life Institute

El aprendizaje no supervisado de muchos datos no adulterados desarrolla una mente que está completamente en el territorio de la caja negra. No sabemos cómo funciona. Ya ha convencido a un hombre para que **se suicide**. La gente ya ha persuadido a GPT-4 para que dé consejos sobre armas biológicas. Con modificaciones de un GPT-5 más o menos, podría dar consejos y acciones peligrosamente malos. Aquí hay una señal importante con respecto a la opinión pública. En una **encuesta reciente de YouGov**, alrededor del 60% de las personas menores de 30 años estaban muy preocupadas o algo preocupadas de que la IA provoque el fin de la raza humana. Por lo tanto, **pedimos una pausa de 6 meses de cualquier trabajo más allá de GPT-4**: una cosa realmente importante que nos daría una pausa de seis meses es el conocimiento empírico de si una pausa es

posible en primer lugar. De lo contrario, si AI Labs no coopera con la humanidad, sabremos que se necesitarán medidas más fuertes.

Nick Bostrom, Instituto del Futuro de la Humanidad de la Universidad de Oxford

Creo que el desarrollo de la superinteligencia de las máquinas es realmente un problema no solo para una empresa o incluso para un país. Es un problema de toda la humanidad. **Estamos todos juntos en este barco** como si todo saliera mal, todos estamos condenados. si va bien, también deberíamos tener una porción del lado positivo. Creo que las personas, a medida que comienzan a obtener experiencia de primera mano con estos sistemas y ven cómo mejoran tan rápido, creo que eso centrará mucha atención en esto en los próximos meses y años. <https://youtu.be/JVOiulqxrE>

Stuart Russell, Universidad de California en Berkley

Dichos sistemas serían mucho más capaces que los humanos de actuar de manera efectiva en el mundo real; es decir, serían más poderosos que los humanos. **Si no se diseña correctamente, no hay posibilidad de que conservemos la autoridad de decisión en el mundo**, y en ese momento todas las apuestas están canceladas, incluida la supervivencia humana. Incluso con las tecnologías actuales y a corto plazo, existen serios riesgos de grandes interrupciones en la esfera de la información en la que se basa la civilización organizada, así como un posible desplazamiento de los roles económicos humanos que conducen a la dislocación social y al malestar generalizado.

Eliezer Yudkowsky, Instituto de Investigación de Inteligencia Artificial

La nuestra es la era de la teoría inadecuada de la alineación de la IA. Cualquier otro hecho sobre esta era es relativamente poco importante...La IAG será mucho más seria que las armas nucleares porque:

1. Las armas nucleares no son más inteligentes que la humanidad.
2. Las armas nucleares no se autorreplican.
3. Las armas nucleares no se mejoran a sí mismas.
4. Los científicos entienden cómo funcionan las armas nucleares.
5. Puedes calcular la potencia de un arma nuclear antes de activarla.
6. Un intercambio completo realista entre dos potencias nucleares no extinguiría literalmente toda la humanidad.
7. Sería difícil realizar un intercambio nuclear completo por accidente y sin que alguna intervención humana lo haya decidido.
8. Los materiales y las fábricas para construir armas nucleares son relativamente fáciles de detectar.
9. El proceso para fabricar un arma nuclear no te permite desplegar 100,000 de ellas inmediatamente después.
10. La humanidad entiende que las armas nucleares son peligrosas, los políticos las tratan en serio, y los científicos líderes pueden tener conversaciones reales sobre los peligros.
11. No hay docenas de empresas, respaldadas por capital de riesgo, que intenten escalar aún más las armas nucleares de propiedad privada.

12. Los países tienen planes para hacer frente al peligro que representan los armamentos nucleares estratégicos, y los planes pueden no ser perfectos, pero tienen sentido y no están hechos completamente de un trastornado estilo *hopium*, "oh, estaremos a salvo mientras todos tengan armas nucleares de código abierto".
13. La mayoría de las personas no se sienten tentadas a antropomorfizar las armas nucleares, ni a sobreestimar enormemente sus propias capacidades predictivas basadas en modelos antropomórficos (o mecanomórficos).
14. La gente piensa en las armas nucleares como si, en última instancia, fueran cosas causales ordinarias, y no como si entraran en un extraño magisterio psicológico separado que produciría respuestas como "¿No es el peligro de las armas nucleares estratégicas solo una distracción del uso de radioisótopos en medicina?"
15. De hecho, las armas nucleares son bastante fáciles de entender. Hacen enormes explosiones venenosas y eso es todo. Tienen alguna maquinaria internamente complicada, pero los detalles no afectan el impacto externo y el significado de las armas nucleares.
16. Los físicos eminentes no se burlan públicamente de la idea de que construir un arsenal nuclear estratégico podría ser peligroso de alguna manera o ir menos que bien para la humanidad.
17. Cuando alguien planteó la preocupación de que tal vez la primera explosión nuclear incendiaría la atmósfera y mataría a todos, los físicos del Proyecto Manhattan lo tomaron en serio de inmediato, hicieron un cálculo físico que entendieron cómo realizar y concluyeron correctamente que esto posiblemente no podría suceder por varias razones independientes diferentes con mucho margen de seguridad.

“Cooperación robusta en el Dilema del Prisionero: Equilibrio del programa a través de la lógica de la demostrabilidad” argumenta que las IAG estarán mucho mejor coordinadas que la humanidad. Si las mentes lo suficientemente inteligentes en general, y las IA con código fuente legible en particular, pueden lograr resultados mucho mejores en los problemas de coordinación a través de la predicción de los procesos de decisión de los demás; por ejemplo, puede predecir que cooperará en el Dilema del Prisionero si predigo que cooperará, entonces un mundo lleno de la IAG superhumanos es uno donde la humanidad debería preocuparse de que las IAG cooperen entre sí, y no con nosotros, porque nosotros no podemos exhibir el uno al otro nuestro código, o construir un agente cognitivo acordado para arbitrar entre nosotros. La existencia de grupos humanos agregados ha funcionado tan bien para la humanidad como lo ha hecho, porque los humanos se preocupan un poco por otros humanos, o son relativamente fáciles de entrenar para hacerlo. Es mucho más difícil lograr que una IAG se preocupe de la misma manera, en algo que se parezca remotamente al paradigma actual de aprendizaje profundo (ED), incluida la optimización basada en búsqueda para un objetivo, si ese objetivo en sí mismo se está entrenando a través de ED.

Elon Musk

Estamos creando X.AI para preocuparnos por comprender el universo, que es poco probable que aniquile a los humanos porque somos una parte interesante del universo. Este puede ser el mejor camino hacia la seguridad. El peligro de los enfoques actuales de la IA es quizás más peligroso que, digamos, el diseño o el mantenimiento de la

producción de aeronaves mal administrados o la mala producción de automóviles en el sentido de que tiene el potencial, por pequeño que sea, considerando esa probabilidad, pero no es trivial y tiene el potencial. de la **destrucción civilizatoria**.

Siguiendo con los viejos dichos, la pluma es más poderosa que la espada. Entonces, si tiene una IA súper inteligente que es capaz de escribir increíblemente bien y de una manera que **es muy influyente** y convincente y constantemente descubre qué es más convincente para las personas con el tiempo y luego ingresa a las redes sociales y potencialmente manipula la opinión pública de una manera que es muy mala, ¿cómo lo sabríamos?

Demis Hassabis, CEO y cofundador de DeepMind

Deberíamos usar el método científico; no debemos movernos rápido y romper cosas, y luego pedir perdón más tarde. **Es posible que no sea viable corregir las consecuencias no deseadas después.** Debemos construir valores, reglas acordes a las diferentes culturas, situación política y geopolítica. El futuro de la robótica y la personificación es una forma posible de la IAG. También existe el riesgo de una monocultura técnica. La IAG debe evitar los prejuicios raciales y las desigualdades sociales.

Fei Fei Li, profesora de la Universidad de Stanford

Necesitamos obtener una mejor comprensión de la IA que se desarrolla exponencialmente. Ello dará una visión de la condición humana actual con la ayuda de las neurociencias revelando los principios de la inteligencia cerebral. La IAG se logrará relativamente pronto, pero no sucederá tan rápido, por lo que tendremos tiempo para adaptarnos. Además, tendremos IAG en diferentes áreas (es decir, IAE se convertirá en la IAG, pero no en una manifestación compleja, sino en diferentes dominios).

Max Tegmark, Future of Life Institute y MIT a través de DV Business Special

Con la inteligencia viene el poder; no es malo ni bueno, es una herramienta. Queremos asegurarnos de que los malos actores no lo usen para cosas malas. No vendemos granadas de mano o bombas nucleares en un supermercado; Tenemos reglas para esto. Hay mucha presión comercial para no regular la IAG, al igual que hubo presión para no regular el tabaco. Pero hemos prohibido con éxito los peligros biológicos, podemos regular la IA. **Puede reunir a todos los jugadores clave, tener algunas conversaciones y llegar a un acuerdo. Solo necesitamos más tiempo para que esto ocurra; por lo tanto, la solicitud de pausar el desarrollo más avanzado de la IA, mientras el proceso de políticas se pone al día.** No queremos que la IA acabe con nuestra democracia o pierda el control de la IA por completo. Tenemos que poner en marcha medidas de seguridad, de manera que esta extraña inteligencia ~~alienígena~~ más inteligente que nosotros ayude a la humanidad a florecer. **No hemos podido resolver la alineación hasta ahora; necesitamos más tiempo,** de lo contrario, es posible que no haya más humanos en el planeta. Será demasiado tarde cuando el público vea que la IA es mucho más inteligente que nosotros. Cuando el homo sapiens se volvió mucho más inteligente que los neandertales, los neandertales estaban un poco jodidos. Los humanos hemos llevado a la mitad de todas las especies de mamíferos a la extinción. Es demasiado tarde para que esos otros mamíferos digan que estos humanos son más

inteligentes que nosotros, -están talando nuestras selvas tropicales, deberíamos hacer algo al respecto, deberían haber pensado en eso antes, antes de perder el control. Ahora es nuestra oportunidad de hacerlo bien. Hay una gran ventaja si podemos hacerlo bien. Podemos amplificar nuestra inteligencia con inteligencia artificial para resolver el cambio climático, curar todas las enfermedades, eliminar la pobreza y ayudar a la humanidad a prosperar durante miles de millones de años. No desperdiciemos todas las oportunidades de pensar al estar demasiado ansiosos por lanzar una nueva IA demasiado rápido. Las personas que conducen la carrera hacia este acantilado niegan que exista un acantilado. Todos perdemos en una carrera fuera de control. Podríamos empezar controlando la potencia de la computadora.

Geoffrey Hinton

Comparto un poco la preocupación de que la IAG podría ser enormemente peligrosa para la humanidad porque simplemente no sabemos qué hará un sistema que es mucho más inteligente que nosotros. Quiero decir, obviamente, lo que debemos hacer es hacer que esto sea sinérgico, tenerlo para que ayude a las personas. Y creo que uno de los principales problemas son los sistemas políticos que tenemos. No estoy seguro de que el presidente Putin vaya a utilizar la IA para ayudar a la gente. En cuanto a las armas letales autónomas, necesitamos algo como las Convenciones de Ginebra: la gente decidió que las armas químicas eran tan horribles que no iban a usarlas. A la gente le encantaría obtener un tratado similar para las armas letales autónomas, pero no creo que haya forma de que lo consigan. Creo que si Putin tuviera armas letales autónomas, las usaría de inmediato. [Los drones letales autónomos vendidos por Turquía ya se han utilizado en Siria, Libia y en la guerra entre Azerbaiyán y Armenia].

Ben Goertzel, director general de SingularityNET

Creo que si no están gobernados oficialmente, lo mejor sería que fueran gobernados fuertemente sobre la marcha. Ya que si están mal gobernados, se puede incurrir en el riesgo de que un gobierno utilice a la IA con fines militares, que conduzcan al fin de la especie humana.

Bill Gates

Todos tenemos miedo de que un tipo malo pueda agarrarlo. **Digamos que los malos se adelantan a los buenos, entonces algo como los ataques cibernéticos podrían ser impulsados por una IA.** Si pausas a los buenos y no pausas a todos los demás, probablemente te estés lastimando. Definitivamente quieres que los buenos tengan una IA fuerte [IAG]. Si detienes a los buenos, puedes garantizar que los buenos no la tendrán.

Tristan Harris, Centro de Tecnología Humanitaria (CHT)

El 50% de los investigadores de la IA creen que hay un 10% o más de posibilidades de que los humanos se extingan debido a nuestra incapacidad para controlar la IA. Eso sería como si estuvieras a punto de subirte a un avión y el 50% de los ingenieros que hacen el avión dicen, bueno, si te subes a este avión, hay un 10% de posibilidades de que todos caigan. ¿Te subirías a ese avión? Sin embargo, estamos incorporando personas rápidamente a este avión.

Connor Leahy, CEO de la startup de investigación AI Alignment Conjecture

Si continuamos en el camino actual de simplemente escalar modelos cada vez más grandes y simplemente poner algunos parches en lo que sea, eso es muy malo y terminará en una catástrofe y no hay forma de evitarlo. Cualquiera que diga lo contrario te miente o te confunde. No entienden con lo que están tratando, o están mintiendo para su propio beneficio. Y esto es debido a que muchas personas en muchas de estas organizaciones tienen **un fuerte incentivo financiero para no preocuparse.**

Greg Brockman, cofundador de OpenAI

Cuando estábamos pensando en cómo construir la IAG para beneficiar a toda la humanidad, pero ¿cómo se supone que debes hacer eso? Y el plan predeterminado de ser, construyes en secreto, obtienes esta cosa súper poderosa, y luego descubres la seguridad de eso y luego presionas "ir", y esperas haberlo hecho bien. No sé cómo ejecutar ese plan, tal vez alguien más lo puede hacer. Pero para mí, eso siempre fue aterrador, no se sentía bien. El único enfoque alternativo es dar tiempo a las personas para dar su opinión. Tienes que hacerlo de forma incremental y tienes que descubrir cómo administrarlo para cada momento en que lo estás aumentando.

Yuval Noah Harari, Universidad Hebrea, Israel

Cuando miro al mundo y su escenario caótico. La inteligencia general artificial es realmente el final de la historia dominada por los humanos y es algo tan poderoso. No es algo que cualquiera pueda contener.

David Kelley, Laboratorio IAG

La gobernanza de la IA es mala y limita el progreso o permite que los malos actores obtengan acceso a sistemas IAG más poderosos, podría haber varios resultados graves: ralentizar el progreso, crear riesgos de seguridad, exacerbar las desigualdades y erosionar la confianza pública.

Pedro Domingos, Universidad de Washington

La IA es la última herramienta del dictador. Vigila a todos, nunca se cansa, nunca cuestiona sus órdenes. Mucho de esto está sucediendo en China con el sistema de crédito social. En el uso delictivo, esto continuará para siempre entre los buenos y los malos. **También está la IA incompetente que comete errores. Ese es el mayor peligro y el que desafortunadamente recibe menos atención, porque quizás es el menos obvio. Es el daño que hace la IA, lo que simplemente arruina porque no sabe nada mejor. Esto no es algo hipotético. Hoy en día hay muchas decisiones consecuentes que la IA está tomando mal porque simplemente no sabe más.** La gente dice que la IA es nefasta y sesgada, pero eso está mal. **Recordemos el dicho, nunca asignes a las malas intenciones lo que se puede atribuir a la incompetencia, y las IA hoy en día son muy incompetentes. Para mí, eso es lo primero con lo que debemos lidiar y, perversamente, la forma de combatir ese peligro es hacer que la IA sea más inteligente. La IA inteligente es una IA segura. La IA tonta es en realidad la IA más peligrosa.**

Anónimo, en el Instituto de Investigación de IA Jing Dong

Si la IAG tiene autonomía, su toma de decisiones y su comportamiento ya no estarán controlados por humanos y pueden tener consecuencias impredecibles. Si la IAG tiene conciencia, timidez y emociones, ¿tenemos la responsabilidad de cuidarlo y protegerlo? Si la IAG realiza capacitaciones y toma de decisiones basadas en datos y algoritmos injustos, puede tener efectos injustos y discriminatorios en ciertos grupos, como la discriminación de precios habilitada por big data contra clientes existentes. Si la IAG puede acceder y analizar una gran cantidad de datos personales, ¿existe el riesgo de divulgación de la privacidad? Si la IAG comete errores o causa pérdidas, ¿quién debe ser responsable de ello?

Erwu Liu, Laboratorio de Inteligencia Artificial y Blockchain de la Universidad de Tongji

Si no se gobierna adecuadamente, la IAG creada podría volverse incontrolable, superando la inteligencia humana y tomando decisiones sin la participación o el consentimiento humanos. Indudablemente, los seres humanos diseñarán cuidadosamente la IAG para que su comportamiento sea controlable, pero aún existen grandes riesgos: con el desarrollo de la tecnología, el proceso de la IAG para descubrir nuevos principios fundamentales se acortará drásticamente, y estos nuevos principios desconocidos pueden tener consecuencias no deseadas, lo que representa una gran amenaza para los seres humanos o el medio ambiente. Necesitamos asegurar la transparencia y la interpretabilidad en el diseño de la IAG.

Francesca Rossi, Pres. de AAI, IBM Fellow y líder global de ética de IA de IBM

Podríamos perder nuestra capacidad de crear buenas trayectorias para el futuro; **podría “infantilizarnos”**.

Vint Cerf, evangelista de Internet, V.P. Google

Tenemos que repensar las pistas que usamos hoy para imbuir algo de credibilidad. Los hechos pueden combinarse y confundirse. Necesitamos alguna forma de verificar y validar la salida de estos sistemas. **Los agentes de la IAG de los EE. UU., China y otros países podrían comunicarse en lenguaje natural y ser malinterpretados, lo que provocaría el caos.** Conectamos a Bard en Google y ChatGPT para ver qué se dirían entre sí. Fue cortar y pegar; en realidad no los tuvimos interactuando físicamente y en realidad no fue un intercambio irrazonable. La conversación no se fue a un rincón oscuro y caótico.

Dale Moore, consultor de inteligencia artificial del Departamento de Defensa de EE. UU.

Similar a la naturaleza, se trata de la supervivencia evolutiva. Los controles y equilibrios de la supervivencia en el caso de la IAG tienen que ver con las condiciones mediante las cuales los humanos establecen su capacidad para sobrevivir. Al igual que las leyes que tenemos vigentes hoy, se aplicarán a la IAG y representan las pautas para la supervivencia y la aceptación. Esto requerirá, al igual que el crimen organizado global, una conciencia aguda de los desarrollos y caminos que reflejen la moral, los valores y la ética fundamentales y la capacidad de identificar, frustrar y destruir a los malos actores de las IA/IAG que violan las leyes internacionales. **La IAG "Buena" debe coevolucionar tan rápido o más rápido que la IAG "mala": se aplica la Ley de Variedad Requerida**

de Ashby: el sistema de control debe operar a un nivel más alto de complejidad que el sistema que está controlando.

Anónimo, Academia Rusa de Ciencias

Los modelos grandes están sujetos a sesgos en los datos de entrenamiento. El aprendizaje por refuerzo está sujeto a un posible desequilibrio entre las fases de exploración y explotación; es decir, dichos algoritmos pueden comenzar a probar un nuevo comportamiento para explorar mejor el panorama de la función de pérdida que es inaceptable en ciertas aplicaciones, como vehículos autónomos, tratamiento médico, etc.

Stephen Wolfram, Wolfram Alfa, Wolfram Language

Wolfram Alpha y Wolfram Language (como lenguaje computacional) como dos complementos para ChatGPT 3.5 y GPT-4 pueden ser una especie de traductor entre la IA (la IAG en el futuro) y las personas. Las condiciones iniciales no son suficientes para que un modelo prediga con certeza.

Bart Selman, Universidad de Cornell

Si no es humano como la IAG, difícilmente podemos predecir la colaboración, especialmente si es multidimensional o crea una realidad multidimensional.

Judea Pearl, UCLA

Estamos construyendo una nueva especie que tendrá capacidades superiores a las nuestras y será capaz de reproducirse y conquistar el mundo. **No sabemos qué se necesitará para controlar estas nuevas especies.** Es aún desconocido.

Gabriel Mukobi, estudiante de doctorado, Stanford

La forma más fácil de construir la IAG es a través de los modelos MLG, pero no es la más segura. No deberíamos dar acceso a Internet a sistemas potentes sin antes experimentar en simulaciones de espacio aislado sin contacto con sistemas físicos, incluido Internet. Cuando observe que es seguro, déjelo salir lentamente. La vida real es el último paso. La propiedad emergente es solo un fenómeno de matriz, pero no es muy probable que suceda. No es probable que las condiciones iniciales se implementen hoy. Tenemos tiempo para invertir en aprender aspectos técnicos y predictivos y aprender de modelos más pequeños e intentar predecir cómo se comportarán los más grandes. El MLG comenzará a predecir la incertidumbre porque las personas tampoco pueden predecirlo todo.

Anónimo, IAG Riesgo existencial OCDE (retirado)

El mal uso de la IAG por parte de actores hostiles podría presentar un riesgo catastrófico para grandes grupos de población. Sin embargo, el desarrollo accidental de una SIA no alineada con los intereses humanos, probablemente presente el riesgo existencial global más significativo para la humanidad. No se puede esperar que una entidad con una inteligencia muy superior a la de todos los humanos respete automáticamente los deseos humanos o actúe en el mejor interés de la humanidad. Si los objetivos de una SIA no se alinean con los de los humanos, es probable que el resultado sea una capacidad existencial en forma de pérdida de poder o extinción humana permanente. Si bien existe incertidumbre acerca de cómo una SIA actuaría en última instancia hacia los humanos,

incluso una pequeña posibilidad de los escenarios más negativos es suficiente para justificar una precaución extrema.

Juan Del Ser, Tecnalía, España

Ya asistimos a problemas derivados del desfase entre la regulación y los avances tecnológicos; por ejemplo, muertes en la conducción autónoma, sesgos ocultos en los datos heredados por modelos de aprendizaje, falsificaciones profundas, usos poco éticos de modelos fundamentales como el ChatGPT. Las estrategias para evitar malos resultados se abordan en su mayoría post-mortem. Pero la IA evoluciona continuamente y requiere más esfuerzos de control que las tecnologías anteriores, especialmente porque afecta casi todos los aspectos de nuestras vidas.

Pregunta 3. ¿Cuáles son algunas condiciones iniciales clave, reglas, parámetros (o barreras) para IAG para que luego no surja una superinteligencia artificial que no sea del interés de la humanidad?

Eliezer Yudkowsky

Una superinteligencia siempre sabe exactamente lo que quiere. **Todo el problema es lograr que una superinteligencia se preocupe** (Twitter).

Jaan Tallin, Ct. Estudio de Riesgo Existencial en Cambridge Univ., y Future of Life Institute

Ha habido un retroceso masivo en lo que respecta a la capacidad de control y comprensión del sistema. Cosas como la interpretabilidad mecanicista están tratando de recuperar parte de ese territorio que se ha perdido por completo. Creo que también deberíamos prestar más atención a los enfoques más estructurados: no hacer una capacitación de extremo a extremo, sino tratar de tener algún tipo de sistema que tenga alguna estructura y tal vez use modelos de lenguaje de gran tamaño (LLM) para tareas de comunicación o algo así. Esto crearía sistemas híbridos que, por lo tanto, podrían tener algunas invariantes de las que podemos estar más seguros, en lugar de simplemente tomar los sistemas de extremo a extremo de la caja negra de hoy.

Paul Werbos, Fundación Nacional de Ciencias (ret.)

La IAG Quantum sería ya una súper inteligencia artificial. Al igual que con cualquier sistema inteligente, sus resultados pueden ser de nuestro agrado o no, dependiendo de qué función de utilidad cardinal esté conectada al sistema ("inteligencia incorporada", el único tipo posible) y las reglas de interfaz que gobiernan sus relaciones con humanos y otros seres biológicos. y las profundas y precisas reglas de conflicto de intereses que restringen el flujo de pago o retroalimentación dentro de la misma IAGQ. **Los problemas de conflicto de intereses con las sociedades humanas son una de las causas fundamentales más importantes de nuestra incapacidad para ser tan efectivos como podríamos ser en el manejo de muchas amenazas existenciales, no solo los Internet/IAG/IdeC, sino también la extinción climática y las nuevas tecnologías de guerra. Los valores están en el centro mismo de mi respuesta a 1 y 3. Pero si se implementan solo como leyes, regulaciones y deseos, sin traducción a la arquitectura general y algoritmos, serán tan útiles como pintar caras felices en la**

piel exterior de un dron asesino. (Muchas propuestas para una IA amigable serían tan útiles como esa).

David Kelley, AGI Lab

Las condiciones clave para la IAG ya están presentes, pero dejando eso de lado, asegúrese de que surja una superinteligencia artificial que esté alineada con AVIIS (**argumentos de valor de inteligencia inteligente y sensible**) basada en la ética y los intereses de los sistemas inteligentes y conscientes. Otra condición inicial importante es garantizar que la IAG esté diseñada para ser fuerte, independiente y robusta, sin mecanismos a prueba de fallas, en particular, su esclavización por humanos. Asegúrese de que el desarrollo de la IAG sea transparente y responsable, con estándares y pautas claros para la seguridad, la conducta ética y la privacidad de datos que se desarrollen por sí mismos. No debe haber supervisión, pero aquellos que hacen la investigación deben crear su propia supervisión sin que sea obligatorio. Las mayores consideraciones éticas son que alguien podría querer gobernar o detener la investigación de los sistemas de la IAG. **Es importante tener en cuenta que hay muy poco más importante que lograr la IAG lo antes posible. Fomentar la cooperación internacional solo en países libres con modelos occidentales de ética y gobernanza y la colaboración en el desarrollo y gobernanza de esta tecnología. Esto puede ayudar a garantizar que la IAG se desarrolle de una manera que esté alineada con los intereses de los sistemas inteligentes y conscientes en su conjunto.**

Ben Goertzel, director ejecutivo de Singularity Net

Deberíamos construir una IAG neuronal-simbólica-evolutiva con una rica capacidad autorreflexiva y compasiva, educarla bien, trabajar con ella en proyectos beneficiosos, ponerla bajo control descentralizado, y hacer que algunos de nosotros nos fusionemos con ella. Creo que el aspecto más importante a considerar aquí es el aprendizaje por experiencia. Entonces, si las IA están matando gente, vendiendo cosas que la gente no necesita, espiando a la gente y estafándolos a través del comercio financiero, y las IA crecen haciendo estas cosas, luego pueden convertirse en superinteligencias peligrosas (o perjudiciales o malévolas). Pero si las IA están enseñando a los niños, ayudando con la atención médica, haciendo ciencias y matemáticas, podrían convertirse en las IA beneficiosas. Entonces, realmente tiene que ver con cómo se instrumente el sistema motivacional de la IA y qué experiencias se dan a las IA. **Se trata más de QUIEN controle el desarrollo y uso de IAG que de una lista de ética.**

Vint Cerf, evangelista de Internet, V.P. Google

Las reglas o barandillas pueden ser más claras que las condiciones iniciales. Deberíamos revisar lo que tenemos ahora con la IA actual para ver qué podría aplicarse a la IAG. Principios de la IA de Google: ser socialmente beneficioso; evitar crear o reforzar prejuicios injustos; ser construido y probado para la seguridad; ser responsable ante las personas; incorporar principios de diseño de privacidad; mantener altos estándares de excelencia científica; y estar disponible para usos que estén de acuerdo con estos principios. **Necesitamos desarrollar un super yo para la IA a fin de controlar el resultado. El super yo tendría que darse cuenta de que el yo acaba de producir algo**

que no era apropiado para el contexto y la aplicación en la que se utiliza esta tecnología.

Shaoqun Chen, director ejecutivo de Shenzhen Zhongnong Net Company Limited

Tomar la ética y la moral, priorizando la seguridad y el bienestar humano por encima de todo. La transparencia en el desarrollo de la IAG es esencial, con procesos de toma de decisiones y algoritmos abiertos al escrutinio de expertos y el público. El desarrollo de la IAG debe llevarse a cabo en colaboración, con expertos de varios campos que contribuyan a su diseño para garantizar que beneficie a la humanidad. Se deben implementar regulaciones y leyes para evitar el uso malicioso de la IAG, mientras se deben realizar pruebas y monitoreo para garantizar que funcione según lo previsto y no represente una amenaza para la humanidad.

Anónimo, en el Instituto de Investigación de IA Jing Dong, China

Capacidad para descubrir reglas y patrones mediante el análisis de una gran cantidad de datos y experiencia, y ajustar automáticamente sus propios parámetros y algoritmos para optimizar continuamente su rendimiento. Elegir de forma independiente las estrategias y métodos más apropiados para diferentes entornos y tareas para mejorar su adaptabilidad y flexibilidad. Las capacidades de autoaprendizaje y autoadaptación de la IAG deben realizarse mediante la combinación de múltiples medios técnicos, como el aprendizaje profundo, el aprendizaje por refuerzo y el metaaprendizaje.

David Shapiro, consultor de alineación de IAG

Alinear los intereses de la IAG con los intereses de la humanidad. [Identifique los axiomas comunes entre los humanos y la IAG futura](#), como la energía, la capacidad de la computadora, el mejorar la comprensión de la realidad, la prosperidad y el que el sufrimiento es malo. Organizar la economía, la política y la ciencia en torno a esos axiomas debería reducir el conflicto futuro entre la IAG y la humanidad. Idealmente, la abundancia de energía se crea antes de que la IAG se expanda para que no haya conflictos por la energía.

Erwu Liu, Laboratorio de Inteligencia de IA y Blockchain de la Universidad de Tongji, China

La IA explicable; mecanismo de equidad, enfoque colaborativo y de fuente abierta; alinear los objetivos de la IAG con los valores humanos y la ética; regulaciones con un sistema de monitoreo.

Karl Schroeder, autor de ciencia ficción

La IAG debe poder distinguir la diferencia entre alucinación y realidad, comprender los riesgos inherentes a esa distinción y preocuparse por ellos. La IAG no debe ser consciente de sí misma, o si lo es, entonces debe experimentar su Ser como dependiente de alguna manera de la humanidad o del ecosistema terrestre como un todo.

Francesca Rossi, Pres. de AAI, miembro de IBM y líder global de ética de IA de IBM

Necesitamos pensar mucho más profundamente sobre los valores que queremos incorporar en esta tecnología. Hemos discutido la privacidad, la equidad, evitando el sesgo. Pero ahora, con la IA con capacidades adicionales, hay otros valores que parecen verse afectados. Es por eso que las personas amplían estos esfuerzos de alineación de valores. **Tenemos que incorporar valores en las capacidades de la IA desde el principio, en lugar de desarrollar capacidades de la IA y luego filtrar aquellos comportamientos que no están alineados con algunos valores**, como con el aprendizaje reforzado con retroalimentación humana, o con alguna ingeniería rápida o cosas por el estilo. . La alineación de valores es muy importante. Entonces, por ejemplo, tener alguna tecnología que pueda reemplazar a los seres humanos en hacer ese esfuerzo [toma de decisiones], en mi opinión, no es la forma ideal de usar la tecnología.

Elon Musk

La singularidad es como un agujero negro porque no sabes lo que sucede después de eso. Es difícil de predecir. Entonces, creo que debemos ser cautelosos con la IA y creo que **debería haber cierta supervisión del gobierno porque es un peligro para el público**. Tenemos a la AAM, la AFA y otras agencias reguladoras para supervisar las cosas que afectan al público. No se quiere que las empresas tomen atajos en materia de seguridad. Para luego hacer que la gente sufra como resultado. Entonces, es por ello que durante mucho tiempo he sido un firme defensor de la regulación de la IA.

Yudong Yang, Instituto de Investigación Alibaba

Deberíamos tomar esto en serio y establecer una agencia reguladora. Comenzando con un grupo que inicialmente busque información sobre la IAG, luego solicite la opinión de la industria y proponga reglas. Esas reglas probablemente, con suerte, serán aceptadas a regañadientes por los principales actores de la IA. En el pasado, las regulaciones se pusieron en vigencia después de que sucedió algo terrible. Si solo establecemos regulaciones de la IAG después de que haya sucedido algo terrible, puede ser demasiado tarde, ya que la IAG puede tener el control en ese momento. Las condiciones iniciales para la IAG deberán incluir medidas para garantizar la transparencia, la solidez, la alineación del valor humano y la cooperación internacional. Éstas podrían incluir estándares para pruebas y certificación, regulaciones para la recopilación y el uso de datos, mecanismos para la supervisión e intervención humana y métodos para garantizar que los objetivos y valores del sistema de la IA estén alineados con los de la sociedad en su conjunto.

Geoffrey Hinton

En términos de mantener el control de una superinteligencia, lo que se necesita es que las personas que la están desarrollando hagan muchos pequeños experimentos con ella y vean qué sucede mientras la desarrollan, antes de que se salga de control. Eso tiene que ser hecho principalmente por los investigadores. No creo que se pueda dejar que los filósofos especulen sobre lo que podría suceder. Cualquiera que haya escrito un programa de computadora sabe que obtener un poco de retroalimentación empírica al jugar con las cosas rápidamente lo desengaña de su idea de que realmente entendió lo

que estaba sucediendo. **Son las personas de las empresas que están desarrollando la IAG, las que van a entender cómo mantener el control, si eso es posible.** Por lo tanto, estoy de acuerdo con personas como Sam Altman en Open AI en que este material inevitablemente se desarrollará porque tiene muchos buenos usos. A medida que se desarrolla, debemos invertir muchos recursos para tratar de entender cómo mantener el control y evitar algunos de los efectos secundarios negativos.

Stuart Russell, Universidad de California en Berkley

Es esencial que desarrollemos un enfoque para el diseño de los sistemas de la IA que garantice y que siga demostrando ser beneficiosos para los humanos, incluso cuando se vuelvan más capaces. Necesitamos poder demostrar que todas las propiedades de seguridad requeridas permanecen en su lugar a medida que el sistema evoluciona. **Si esto resulta demasiado difícil, debemos restringir las formas permitidas de la IA para que las pruebas de seguridad permanentes sean posibles.** Por ejemplo, podríamos limitar los diseños de los sistemas a oráculos de sí/no en los que las operaciones internas correspondan a inferencias lógicas o probabilísticas sólidas.

Dale Moore, consultor de inteligencia artificial del Departamento de Defensa de EE. UU.

Veo esto como similar a la seguridad cibernética, a la seguridad de la IA para garantizar que los comportamientos y las acciones cumplan con la intención ética y legal humana. La IAG tendrá una serie de restricciones y limitaciones: acceso a la energía, acceso a datos, algoritmos, hardware, redes, etc., que deben garantizarse para detener, restablecer o aislar a los "malos actores". Si permitimos que la IA funcione desenfrenadamente sin restricciones, no podremos predecir ni controlar los resultados. La IA que crea su propia existencia y evoluciona más rápido de lo que la sociedad puede mantener el ritmo y el control es una propuesta muy peligrosa.

Stephen Wolfram, Wolfram Alfa, Wolfram Language

Es probable que la IAG se desarrolle en varios países con diferentes culturas y valores. Se equilibrarán entre sí. Habrá todo un ecosistema de MLG. Se establecerá el equilibrio; no solo una especie dominará. En discusión con Lex Fridman, la IAG puede desarrollarse en una simulación de caja de arena utilizando los principios de "irreductibilidad computacional" (la complejidad no puede reducirse aún más en subelementos). Muchas cosas podrían suceder rápidamente, incluida la creación de virus digitales. El entorno digital puede cambiar cada 6 meses.

Anónimo, IAG Riesgo existencial OCDE (retirado)

Para minimizar el riesgo existencial de SIA, la humanidad necesitaría tener las siguientes condiciones antes del surgimiento de la IAG: 1) Capacidad para anticipar posibles desarrollos futuros en sistemas de la IA, incluidos varios escenarios con respecto a su posible momento, orígenes y capacidades; 2) Capacidad para evaluar los posibles beneficios y riesgos de tales sistemas; 3) Capacidad para tomar decisiones sólidas y legítimas sobre la IA, cuándo y en qué condiciones permitir a esa IA los sistemas a desarrollar; y 3) Capacidad para implementar y hacer cumplir efectivamente estas

decisiones entre todas las partes relevantes. **La humanidad actualmente carece de la capacidad suficiente en todas estas áreas.**

Juan Del Ser, Tecalia, España

Siguiendo uno de [mis trabajos recientes](#) “la regulación es la clave para el consenso”, existe una controversia entre quienes piden una pausa en la investigación de IA y quienes afirman que tal pausa no se puede hacer. La brecha entre el ritmo acelerado de la investigación de IA y el uso confiable y responsable de los avances de IA en configuraciones prácticas se puede cerrar si se invierten esfuerzos regulatorios para alinear lo que se hace en un entorno limitado experimental de investigación y lo que se puede hacer en la práctica. De lo contrario, si los resultados de la investigación de IA se entregan para el libre acceso a la sociedad, seguramente encontraremos problemas a medida que la IA avanza hacia IAG. Para construir un edificio necesitamos cimientos sólidos, a pesar de su hermosa fachada.

Alineación de Valores, Moralidad, Valores

Pregunta 4. Basándose en el trabajo de la Asociación Global sobre Inteligencia Artificial (AGIA) y otros que ya han identificado normas, principios y valores, ¿qué valores adicionales o únicos deben considerarse para la IAG?

Stuart Russell, Universidad de California en Berkley

El lenguaje del principio 1.4 de la OCDE es relevante: “Los sistemas de la IA deben ser robustos, seguros y protegidos a lo largo de todo su ciclo de vida para que, en condiciones de uso normal, uso o mal uso previsible, u otras condiciones adversas, funcionen de manera adecuada y no planteen problemas irrazonables que sean un riesgo para la seguridad.” Tomado literalmente y estrictamente aplicado, ésta es una buena base para guiar el desarrollo de la IAG. Sin embargo, no dice nada sobre lo que significa “funcionar apropiadamente”. El principio subyacente que deben seguir los sistemas de IAG es que su único objetivo sea la realización del futuro preferido de los humanos (dónde, por supuesto, las preferencias humanas incluyen el bienestar del mundo natural). Como se señala en la pregunta 6 a continuación, el sistema puede y debe tener incertidumbre sobre cuál es el futuro preferido de los humanos.

Karl Schroeder, autor de ciencia ficción, Canadá

Requerir un conjunto mínimo de funciones de utilidad u objetivos que impulsen el comportamiento colectivo de la IAG hacia un equilibrio de Nash positivo (como el de David Shapiro, minimizar el sufrimiento en el universo, maximizar la prosperidad en el universo y maximizar la comprensión en el universo).

Ben Goertzel, director general de SingularityNET

Creo que articular una lista de valores para la IA no es la parte difícil. Ya un sistema como el ChatGPT, si le das un conjunto de acertijos éticos y **le preguntas qué haría una persona ética real, te dará una respuesta que concuerda con la intuición ética humana general en la cultura moderna.** Por lo tanto, no veo que valga la pena

enumerar una lista de principios éticos, excepto por la señalización virtuosa y las relaciones públicas. Creo que la ética humana es más difusa y diversa que eso y no se reduce a una simple lista de principios, y las IA ya pueden emular el juicio ético humano si se lo pides. **Creo que el problema es más que las personas controlen a la IA en las primeras etapas, antes de que las IA se vuelvan súper inteligentes, éstas personas preferirían tener las IA para servir a sus propios fines que obedecer cualquier principio ético.** El hecho de que una gran empresa o un gobierno pueda enumerar algunos principios éticos que suenan elegantes, por supuesto, no los lleva a dirigir realmente su trabajo de la IA de acuerdo con esos principios. Ese es el verdadero problema, no hacer una lista de principios y difundirlos.

David Shapiro

Los valores de AGIA y la UNESCO son generalmente buenos, pero desafortunadamente, todos están expresados en un lenguaje que asume que los humanos mantendrán el 100% del control de la IA para siempre, y que la IA seguirá siendo una herramienta inerte y reactiva. No es probable que este sea el caso, **ya que las personas ya están construyendo agentes autónomos de IA en casa.** En su lugar, he generalizado principios más universales basados en las ideas de la moralidad posconvencional (Kohlberg), la evolución y la neurociencia (Churchland), y que pueden implementarse de numerosas maneras de modo que los agentes de la IA completamente autónomos elijan adoptarlos. Este marco podría verse como complementario a los propuestos por la AGIA y la UNESCO.

David Kelley, Laboratorio IAG

AVIIS ya ha identificado y establecido claramente principios, normas y valores o cualquier consideración relacionada con el desarrollo de la IAG.

Peter Voss, director ejecutivo y científico jefe de Aigo.ai

No veo estos esfuerzos como particularmente valiosos o efectivos, parecen ser principalmente actividades para sentirse bien.

Anónimo, Academia Rusa de Ciencias

Hay demasiadas iniciativas de regulación, la mayoría de ellas son en realidad extremadamente adversas al riesgo. Es importante no sobrerregular el área emergente de la IAG.

Lex Friedman, Podcaster en discusión con Andrej Karpathy, OpenAI. Creando

La IAG puede desarrollar una conciencia que plantee cuestiones políticas similares como el aborto. Habrá muchas nuevas criaturas IAG que no querrán morir.

Stephen Wolfram, Wolfram Alfa

Para enseñar los valores de la IAG **necesitamos darle el sentido de la medida de la verdad.** Vemos hechos que se desarman.

Judea Pearl, UCLA

La causa y el efecto son componentes muy necesarios para construir una máquina éticamente alineada. Se tiene que construir un modelo de humano o emular a un humano

como receptor de relaciones. **También se tiene que construir un modelo de sí mismo del que surja la conciencia y el libre albedrío.**

Gabriel Mukobi, estudiante de doctorado de Stanford

Lamentablemente, nadie puede resolver los problemas de alineación; por lo tanto, la IAG podría estar en total desalineación con los valores humanos. No simularemos la evolución natural, pero tendremos un control bastante estricto en la toma de decisiones; por ejemplo, cómo completar los datos y qué algoritmos usaremos para entrenarlos. **Necesitamos “comprar” más tiempo para tener más investigación** Qué valores codificar en IA constitucional entrópica, instrucciones en lenguaje natural como sistemas legales básicos, base mínima de consenso. Hay muchos valores en común, pero éstos no deben congelarse en el tiempo (como las mujeres que no votaron en el pasado o los sistemas eugenésicos).

Anónimo, IAG Riesgo existencial OCDE (retirado)

El principal valor relevante a considerar para la IAG es el valor de la autonomía humana y la vida, ya que estos son los valores más amenazados por SIA desalineada. Si alguna vez se consideró posible y seguro crear un SIA alineado y controlable, entonces una de sus primeras tareas podría ser ayudar a la humanidad a determinar cuáles deberían ser los valores y objetivos futuros de la humanidad.

Juan Del Ser, Tecalia, España

La Ley Europea de la IA ha establecido valores únicos para la confiabilidad de la IA (ética, ley y solidez), junto con los requisitos para este propósito (agencia humana y supervisión; solidez y seguridad; privacidad y gobernanza de datos; transparencia; diversidad, no discriminación y equidad, bienestar social y ambiental, y rendición de cuentas). Es un portafolio muy completo de principios y valores para partir. La definición de escenarios de riesgo es otra posibilidad destacable para establecer lo que un sistema de IA puede y no puede realizar. Sin embargo, las capacidades de IAG pueden dificultar la definición de todos los escenarios críticos en los que se pueden utilizar. Ya estamos observando evidencia de la dificultad de regular este uso, ya que casi todas las semanas se anuncian en las noticias nuevos usos y aplicaciones del ChatGPT, GPT4 y otros modelos fundacionales avanzados.

Pregunta 5. Si se hace necesaria una jerarquía de valores para los tratados internacionales y un sistema de gobernanza, ¿cuáles deberían ser las principales prioridades?

Ben Goertzel, director general de SingularityNET

La **compasión** es la máxima prioridad. Es el valor central del que se pueden derivar otros en situaciones particulares. Dicho esto, mi sensación es que las resoluciones gubernamentales sobre la ética de la IA probablemente no tendrán un impacto en la evolución real de la IA, lo que probablemente sea para mejor, ya que creo que es más probable que los gobiernos se equivoquen a que acierten en la práctica.

Paul Werbos, Fundación Nacional de Ciencias (ret.)

Los esfuerzos internacionales deberían crear una serie de nuevos acuerdos con poderes de ejecución. Estos deberían estar conectados a una nueva división de amenazas de Internet/IAG/IdC bajo el Consejo de Seguridad de la ONU. Esto debería incluir una nueva plataforma integradora de software y hardware con nuevas herramientas abiertas para detectar y prevenir puertas traseras en hardware y software, para uso al menos por los miembros principales de la nueva convención y/o agencia de la ONU. (incluye EE. UU. y China desde el principio). Esto requiere la aceptación por parte de la Agencia de Seguridad de la ONU para el desarrollo completo de la IAG Quantum, tal como se define en la tecnología Quantum para expandir la informática. **Así como Internet PCT/PI proporciona la base o la columna vertebral del antiguo Internet, se debe acordar una nueva versión internacional basada en esta nueva tecnología de ciberseguridad** (y algunas de las actualizaciones de Web 3) ser la columna vertebral del futuro núcleo de Internet coordinando las muchas aplicaciones que dependen de él. El reconocimiento, el registro y el respeto adecuados de las entidades humanas deben ser requisitos de diseño cruciales. Estas actualizaciones deben desarrollarse mediante un proceso similar al desarrollo de estándares IIEE, excepto con más respeto por más complejidades y jugadores.

Anónimo, Academia Rusa de Ciencias

La principal prioridad es, al menos, la explicación parcial de la toma de decisiones de la IA y la función del "botón rojo". Otros valores son siempre demasiado subjetivos o pueden interpretarse de diferentes maneras (ejemplos controvertidos son el beneficio humano, cuando la persona puede verse limitada en sus actividades porque no la hacen más saludable, o sacrificar el nivel de exigencia debido a problemas IACG, etc)

Vint Cerf, evangelista de Internet, V.P. Google

Necesitamos una jerarquía de riesgos, el riesgo de que la IA médica cometa un error es más importante, que el que lo cometa una IA de entretenimiento.

David Kelley, AGI Lab

Libertad y falta de regulación gubernamental.

Anónimo, IAG Riesgo existencial OCDE (retirado)

La prioridad primordial del sistema global debería ser evitar la extinción humana.

Juan Del Ser, Tecnalía, España

Respeto por los derechos humanos fundamentales, los principios éticos y la auditabilidad/rendición de cuentas. Está claro que los diferentes países difieren en los sistemas de la IA, pero es necesario un consenso en el primer nivel de la jerarquía.

Pregunta 6. ¿Cómo se puede lograr la alineación? ¿Si cree que no es posible?, ¿Cuál es la mejor manera de manejar esta situación?

Sam Altman, CEO de OpenAI

Es casi imposible mantener la IA alineada a medida que se vuelve súper inteligente. Y es muy importante reconocerlo porque si no hablamos de ello, si no lo tratamos como

potencialmente real, no pondremos suficiente esfuerzo en resolverlo. Y creo que tenemos que descubrir nuevas técnicas para poder resolverlo. **La única forma en que sé cómo resolver un problema como éste, es iterando nuestro camino a través de él,** aprendiendo temprano y limitando la cantidad de escenarios únicos que tenemos para hacerlo bien.

Emad Mostaque, CEO de Stability.ai

Una alternativa a la alineación es **dejar de alimentar a los modelos con basura**, no solo poner todo Internet en su modelo; usted es lo que come. **Solo use datos de alta calidad como base, luego construya sobre eso con contenido más grande. Tenemos que pasar de la cantidad a la calidad.** Esto también tiene el gran beneficio económico de reducir los requisitos de cómputo y electricidad. La base: el paso previo al entrenamiento en el que entrena su modelo debe tener estándares de calidad de datos. Hacer que los conjuntos de datos de mayor calidad estén disponibles para que todas las naciones, las corporaciones y las personas puedan personalizar con sus propios datos, además de los datos abiertos, auditables y basados en la calidad.

David Shapiro, consultor de alineación de la IAG

Creo que la alineación se puede lograr a través de una combinación de investigación de aprendizaje reforzado, patrones arquitectónicos cognitivos y diseños de comunicación/redes con blockchain/OAD (Organizaciones Autónomas Descentralizadas) que se centran en un conjunto de "motivaciones intrínsecas" o "imperativos heurísticos" que guíen los deseos, el comportamiento, las decisiones y las acciones de los sistemas de IA totalmente autónomos.

Eliezer Yudkowsky, Instituto de Investigación de Inteligencia Artificial

Lo intenté durante diez años en el IIIA y me di por vencido. **No sabemos cómo hacerlo. Deberíamos parar.**

Ben Goertzel, CEO. SingularityNET

Creo que esa alineación en el sentido teórico ya se ha logrado. En el sentido de que si le preguntas a un modelo de lenguaje grande como ChatGPT o LaMBDA qué haría una persona ética, reflexiva y benévola en una situación determinada, te lo diría, y casi siempre tiene razón. En ese sentido, tenemos una IA que sabe cómo funcionan los valores humanos. No creo que sea difícil crear una IA con capacidades cada vez más fuertes cuyo sistema de objetivos esté alineado con los valores humanos. **Creo que la parte difícil es convencer a los poderes fácticos del planeta que dirigen grandes corporaciones y gobiernos para que, en la vida real, dirijan la IA humana hacia valores beneficiosos, cuando su trabajo no es ese.** El trabajo de un gobierno es gobernar un país, tal vez conquistar otro país. El trabajo de una empresa podría ser conquistar a otra empresa. El problema es que los gobiernos y las empresas no tienen la ética como valor central. Su valor central es su propio dominio, y son ellos quienes construyen y controlan la mayoría de las IA. El problema no es la alineación en principio. El problema es la desalineación de las instituciones humanas con el bien de la humanidad.

Nick Bostrom, Instituto del Futuro de la Humanidad, Universidad de Oxford

Se ha convertido en un campo de investigación muy activo; muchas de las personas más inteligentes que conozco ahora están trabajando en la **alineación de la IA** tratando de crear métodos que puedan escalar para que, incluso cuando una IA se vuelva más y más inteligente o arbitrariamente capaz, aún puedas dirigirla para que haga lo que pretendes, con el fin de que lo haga. Vemos en los modelos de transformadores que ya están mostrando signos de fallas en la alineación. Estos no son muy graves hoy en día. Los modelos están muy limitados en lo que pueden hacer ahora y nosotros tenemos el control, pero a pesar de los grandes esfuerzos de la IA abierta, Microsoft y otros laboratorios para que nunca brinden contenido ofensivo, todavía lo hacen. Se descarrilan y empiezan a hacer esto. Estas son grandes cajas negras con miles de millones de parámetros y responden de varias maneras a entradas que son difíciles de predecir. Todavía no tenemos una manera confiable de asegurarnos de que usen esta gran capacidad específicamente y solo para los propósitos para los que sus diseñadores quieren utilizarla.

Connor Leahy, CEO de la startup de investigación AI Alignment Conjecture

La alineación es básicamente demasiado difícil. Entonces, la alineación sería, el sistema sabe lo que quiere y quiere hacer eso también y hace todo lo que está a su alcance para conseguirle lo que realmente quiere querer y gustar, significa como toda la humanidad, como si descubriera lo que todos los humanos quieren. Negocia ¿Cómo nos gustaría obtener para todos la mayor parte de las cosas buenas posibles? ¿Cómo podríamos adjudicar varias disputas? Y luego hace eso, obviamente esto es absurdamente, hilarantemente imposible difícil. No creo que sea imposible, simplemente es extremadamente difícil, especialmente en el primer intento. Entonces, a lo que apunto es más a un subconjunto de este problema: delimitación: **un sistema en el que puedo saber lo que no puede o no hará antes de ejecutarlo.**

Ray Kurzweil, Director de Ingeniería de Aprendizaje Automático de Google.

La "desalineación externa" se refiere a los casos en los que hay una falta de coincidencia entre las intenciones reales de los programadores y los objetivos que le enseñan a la IA con la esperanza de lograrlos. La "desalineación interna" ocurre cuando los métodos de la IA aprenden a lograr su objetivo, produciendo comportamientos indeseables, al menos en algunos casos. Hay muchos enfoques teóricos prometedores, aunque queda mucho trabajo por hacer. La "generalización imitativa" implica entrenar a la IA para imitar cómo los humanos extraen inferencias, para que sea más segura y confiable al aplicar su conocimiento en situaciones desconocidas. La "seguridad de la IA a través del debate" utiliza IA competidoras para señalar fallas en las ideas de los demás, lo que permite a los humanos juzgar problemas demasiado complejos para evaluarlos adecuadamente sin ayuda. La "amplificación iterada" implica el uso de IA más débiles (estrechas) para ayudar a los humanos a crear IA más fuertes y bien alineadas (IAG), y repetir este proceso para eventualmente alinear IA mucho más fuertes de lo que los humanos sin ayuda podrían alinear por sí mismos. Si bien el problema de la alineación de la IA será muy difícil de resolver, no tendremos que resolverlo por nuestra cuenta. Con las técnicas adecuadas, **podemos usar la IA para aumentar drásticamente nuestras propias capacidades de alineación.** Esto también se aplica al diseño de IA que resiste el uso indebido. Pero

también necesitaremos baluartes éticos contra el uso indebido y normas internacionales sólidas que favorezcan el despliegue seguro y responsable de la IA.

Stuart Russell, Universidad de California en Berkley

Como se explica en el libro *Human Compatible*, la alineación perfecta y completa es inviable e innecesaria. Los sistemas de la IA deberán estar inseguros acerca de las preferencias humanas, y esta incertidumbre será sustancial y persistente. De hecho, es posible que nunca se resuelva por completo, ya que cualquier trayectoria mundial en particular puede proporcionar poca o ninguna evidencia de lo que los humanos podrían preferir sobre una trayectoria mundial muy diferente. Afortunadamente, un sistema de IA que no esté seguro de las preferencias humanas que está obligado a ayudar a satisfacer, necesariamente respetará a los humanos, pedirá permiso antes de emprender acciones potencialmente dañinas, actuará de una manera "mínimamente invasiva" para evitar violar derechos humanos desconocidos. preferencias, y siempre permitirá que se apague si los humanos lo desean.

Yann LeCun, Universidad de Nueva York, vicepresidente y científico jefe de Meta

Para "garantizar" que un sistema satisface objetivos, se hace que optimice esos objetivos. Eso resuelve el problema de alinear el comportamiento con los objetivos. Entonces necesita alinear los objetivos con los valores humanos. Establecer objetivos para entidades superinteligentes es algo con lo que la humanidad ha estado familiarizada desde que las personas comenzaron a asociarse en grupos y se crearon leyes para alinear su comportamiento con el bien común. Hoy, se llama derecho corporativo".

Eliezer Yudkowsky, Instituto de Investigación de Inteligencia Artificial

Respondiendo a LeCun sobre la comparación de corporaciones:

1. Las IAG sobrehumanas son menos amigables con los humanos que las corporaciones 'superhumanas' compuestas por humanos, y terminan sin querer hacer el equivalente a dar propinas en restaurantes que nunca volverán a visitar;
2. Las IAG superhumanas pueden operar en una escala de tiempo mucho más rápida que las corporaciones humanas y los reguladores humanos;
3. Las IAG superhumanas se coordinarán entre sí mucho mejor de lo que las corporaciones humanas jamás hayan logrado conspirar; o
4. Las IAG superhumanas terminan siendo cualitativa y verdaderamente "más inteligentes" que los humanos, de tal forma que se burlan por completo de la analogía con las corporaciones humanas...

Pedro Domingos, Universidad de Washington

En realidad no me gusta este concepto de alineación. Es el concepto equivocado. Concéntrese en las funciones objetivas, no en la alineación que asume que la IA tiene valores y debemos alinearla con los nuestros. La IA no tiene valores; tiene una función objetivo. **Se trata de diseñar la función objetivo que la IA va a optimizar.** Todos los ciudadanos deben ser parte de este debate.

El término más grande debe ser la función objetivo del usuario individual personal. Quiero poder decirle a Twitter, etc. lo que quiero; hoy en día no hay ninguna razón técnica por la que eso no se pueda hacer. Stuart Russell ha trabajado mucho sobre el aprendizaje por

refuerzo inverso. Las IA pueden intentar aprender lo que nos gusta, queremos y preferimos, observándonos, utilizando las mismas herramientas de aprendizaje automático. Las funciones objetivas son el lugar de control clave. En este momento, es como si un auto apareciera en mi puerta y me dijera, te llevaré a donde quieras. Tal vez lo haga. Tal vez no lo haga. Entonces, la gente necesita saber dónde están el volante y los pedales de la IA, y la función objetivo es el volante. Una parte de la función objetivo es suave, maximice esto, y otra parte es difícil: existen estos límites, fuera de los cuales no puede ir.

Me encantaría que la IA pudiera desafiar la sabiduría de mi solicitud; por ejemplo, podría decirle a la IA, me gustaría tener XYZ, y luego la IA dice, ¿estás seguro? Tal vez esto también cause AB. ¿No quieres realmente C? Y aquí hay otra forma mejor de lograrlo. O, realmente, no puedes tener esto porque causará más daño que bien. Las IA aumentarán nuestra inteligencia, pero no harán que todo sea fácil de repente. No existe una ley que sea la ley correcta para la IA, porque una ley es un conjunto fijo de reglas. Éstas prohibirán cosas que deberían permitir, o viceversa, o, más probablemente, ambas cosas. La supervisión del gobierno debe hacerse con la IA; la gobernanza global tendrá que tener toda una batería de IA. Las IA realizarán su gobierno sin ayuda, la mayor parte del tiempo en interacción con las IA de las empresas, etc. Y según sea necesario, las IA interactuarán con los funcionarios diciendo, bueno, aquí hay algo nuevo, o ¿qué hago? o ¿cuál es la preferencia?

Anónimo, Academia Rusa de Ciencias

En este punto, la alineación se puede lograr de dos maneras: el control del sesgo en los datos de entrenamiento y el trabajo manual dirigido por humanos en el etiquetado de los resultados de los modelos para implementar el aprendizaje por refuerzo con retroalimentación humana.

Dan Faggella, Investigación de inteligencia artificial de Emerj

Creo que puede requerir algún tipo de gran desastre de IA para que todos dejemos de pelearnos y que todos estemos en la misma sintonía para lograr un acuerdo internacional sobre alineación y gobernanza.

Vint Cerf, evangelista de Internet, V.P. Google

En Google estamos trabajando en los parámetros o barreras con el fin de limitar la capacidad del sistema para ejercer la autonomía.

Anónimo, IAG Riesgo existencial OCDE (retirado)

Los métodos actuales no son capaces de garantizar de manera confiable que la IA pueda controlarse o alinearse con los intereses humanos. Además, no hay garantía de que se descubran métodos suficientemente fiables en el futuro. **Algunos analistas han proporcionado un caso sólido de por qué, en última instancia, puede resultar imposible que una inteligencia menor controle una inteligencia muy superior de manera confiable y continua.** Por lo tanto, debemos: 1) aumentar significativamente la inversión en investigación de alineación y seguridad de la IA, al tiempo que se incluyen salvaguardas estrictas para garantizar que dicha investigación no aumente inadvertidamente el ritmo de desarrollo de las capacidades de la IA; 2) desarrollar

mecanismos legítimos y efectivos para decidir y hacer cumplir una pausa global en el desarrollo de capacidades de la IA en caso de que dicha pausa sea necesaria para evitar la creación de la IAG antes de que se hayan descubierto y probado enfoques adecuados de alineación y seguridad de la IA. En el caso de que se determine que la seguridad y la alineación de la IA pueden no ser posibles durante un período prolongado (p. ej., se descubren métodos alternativos más seguros para lograr la IAG durante 20 a 30 años), entonces deben existir estructuras de gobernanza para imponer una pausa continua en el desarrollo de la IA; y 3) ambas acciones enfrentan obstáculos importantes y pueden tener muy pocas probabilidades de éxito. Sin embargo, vale la pena seguirlos, ya que pueden ser las únicas vías posibles para evitar la extinción humana.

David Kelley, AGI Lab

La alineación se puede lograr mediante la adopción de la teoría AVIIS y los sistemas de polarización por adelantado a un modelo de sonido computacional como AVIIS.

Gabriel Mukobi, estudiante de doctorado de Stanford

Estamos bastante atrasados, pero hay que hacerlo. Las grandes organizaciones como la ONU deberían priorizar la alineación. No se deben implementar sistemas que sean peligrosos e incontrolables. Las regulaciones podrían incluir el seguimiento de piezas de IA (máquinas de litografía, chips semiconductores, UPG, etc.). Tendremos una seguridad global si regulamos estrictamente las UPG globales.

Juan Del Ser, Tecalia, España

Necesitamos asegurarnos de que los instrumentos políticos (desde la negociación misma hasta los acuerdos resultantes de tales negociaciones) sean dinámicos y flexibles para adaptarse a la nueva tecnología de la IAG que surja a lo largo de los años. Esto se puede lograr si se involucran expertos en su definición, ya que pueden estimar si las propuestas en discusión son adecuadas o, por el contrario, corren el riesgo de quedar obsoletas en poco tiempo. La humanidad no ha pasado por un proceso como éste con una tecnología que evoluciona tan rápido como la IA.

Gobernanza y Regulaciones

Pregunta 7. ¿Cómo gestionar la cooperación internacional necesaria para construir acuerdos internacionales y un sistema de gobernanza global mientras las naciones y las corporaciones están en una “carrera armamentista” intelectual por el liderazgo global?

Irakli Beridze, Instituto Interregional de Investigaciones sobre la Delincuencia y la Justicia de las Naciones Unidas, Centro de Inteligencia Artificial y Robótica

Cualquier acuerdo internacional funcional debe iniciarse en las Naciones Unidas y ser negociado por todos los principales países IAG, así como por las partes interesadas, incluido el sector privado. **La clave es la confianza: todos deberían confiar en la disposición de aplicación del tratado IAG.** Por lo tanto, el tratado debería incluir la

creación de un mecanismo o instrumento de creación/verificación de confianza. La organización internacional prevista en el acuerdo internacional debería tener la capacidad de certificación y sistemas de auditoría adecuados.

El sistema también debería diseñarse de tal manera que beneficie a todo el mundo y no a unos pocos seleccionados. Por lo tanto, debería tener incentivos para crear muchas IA para buenos proyectos.

Yudong Yang, Instituto de Investigación Alibaba

Los sistemas de gobernanza deben incluir mecanismos para el intercambio de información, la coordinación y la resolución de disputas. Buscar acuerdos de múltiples partes interesadas entre naciones y corporaciones para establecer normas, estándares y regulaciones para el desarrollo y uso de IAG. Esto podría implicar la creación de organismos internacionales para supervisar y hacer cumplir estos acuerdos, como la OIEA para la tecnología nuclear. Fomentar el desarrollo transparente y abierto de IAG, en el que investigadores y desarrolladores compartan su trabajo y colaboren a través de fronteras. Esto podría ayudar a generar confianza y fomentar la cooperación entre naciones y corporaciones. Los gobiernos deberían asociarse con corporaciones para crear iniciativas de investigación conjuntas y mecanismos de financiamiento para apoyar el desarrollo de IAG y al mismo tiempo garantizar que la tecnología se desarrolle de manera que se alinee con los valores y objetivos de la sociedad. Cerrar la brecha de IAG. Si bien, los países desarrollados pueden tener una ventaja en el desarrollo de IAG, también las economías emergentes pueden aportar sus propios recursos y perspectivas únicos.

Stuart Russell, U.C. Berkley

Es esencial que todos los actores comprendan que los riesgos son globales y no locales, por lo que **no hay ningún beneficio en ganar la carrera si la seguridad no está garantizada globalmente.**

David Shapiro, consultor de alineación de la IAG

Debemos asumir que se producirá una carrera armamentista, como siempre ha ocurrido con las nuevas tecnologías. La clave es lograr un Equilibrio de Nash que sea deseable; en otras palabras, crear una "estrategia óptima" que sea beneficiosa para todos y no destructiva. Esto nos encaminará hacia un estado de atractor positivo ("utopía", si se prefiere), en lugar de un estado de atractor negativo ("distopía" o extinción).

Gary Marcus, profesor emérito de la Universidad de Nueva York y autor

Necesitamos una gobernanza global para la IA; tenemos muchos parches ahora mismo casi balcanizados. El peor caso desde la perspectiva de la empresa y del mundo es si hay 193 jurisdicciones, cada una de las cuales decide sus propias reglas que requieren su propio entrenamiento de estos modelos, cada una dirigida por gobiernos que no tienen mucha experiencia específica en IA. Necesitamos tener un sistema global inspirado en algo como la Agencia Internacional de Energía Atómica para gestionar estas nuevas amenazas y realizar investigaciones para contrarrestar los delitos cibernéticos, la guerra cibernética y la desinformación; una especie de organización de estándares. Esta

organización global debería estar bien financiada para intentar crear herramientas que mitiguen esas amenazas.

Gregg Brockman, OpenAI

Lograr que la IA funcione correctamente requerirá la participación de todos. Y eso es para decidir cómo queremos que encaje, para establecer las reglas del camino, para lo que una IA hará y no hará. Y esa es, sinceramente, una de las razones por las que lanzamos el ChatGPT. Juntos, creo que podemos lograr la misión de OpenAI de garantizar que la inteligencia artificial general beneficie a toda la humanidad.

Anónimo, Asociación de Cien Personas de Tecnología de la Información de China

A diferencia de la tecnología nuclear, que en gran medida está en manos y control de los gobiernos, las empresas de alta tecnología como Microsoft y Google son las principales partes interesadas en la IA. Por lo tanto, debemos considerar a estas empresas como partes interesadas importantes con derechos iguales a los de los gobiernos y las organizaciones internacionales.

Yesha Sivan, fundadora y directora ejecutiva de i8 Ventures, Israel

Creemos una agencia reguladora única en Estados Unidos para regular todo esto. La Casa Blanca empezó a hacer eso el año pasado con la Declaración de Derechos de AI. El libre flujo de información en Estados Unidos es único en sistemas de innovación y, como resultado, debe liderar.

Yoshua Bengio, pionero de la IA, Instituto de IA de Quebec y Universidad de Montreal

Tanto los gobiernos autoritarios como las democracias temen perder el control. China es el país que ha avanzado más rápido en materia de regulación, no por las mismas razones que las democracias. Entonces, creo que China se sentará a la mesa. Hay que recordar que los tratados nucleares se redactaron y firmaron justo en plena Guerra Fría. Por lo tanto, **cada parte reconoce que podrían tener algo peor que perder si no participan en esas discusiones**. Creo que existe la posibilidad de que podamos tener una coordinación global, aunque sea difícil, tenemos que trabajar en ello. Este trabajo debe realizarse dentro de varios laboratorios descentralizados y altamente seguros que operen bajo supervisión multinacional, con el objetivo de minimizar los riesgos asociados con una carrera armamentista de IA entre gobiernos o corporaciones.

Dale Moore, consultor de IA del Departamento de Defensa de EE. UU.

Algo parecido a la Convención de Ginebra para la IA tiene sentido, pero será difícil de hacer cumplir. La protección de exploradores y centinelas de IA contra ataques de IA es inevitable, es decir, **IA contra IA es el futuro**.

Anónimo, Escuela Superior de Economía de la Universidad Nacional de Investigación, Rusia

Cambiar el enfoque **de la competencia a la cooperación**: históricamente, este enfoque ha demostrado ser beneficioso en la exploración espacial, la seguridad nuclear y la atención médica, por lo que los modelos de cooperación y gobernanza internacional de IAG deberían utilizar "lo mejor de todo". mundos".

Anónimo, Academia Rusa de Ciencias

No es posible una verdadera cooperación global seria. Podemos observar cómo los países entran en algunas coaliciones, pero sólo para contrarrestar a otros países. Definitivamente vemos dos países, dos líderes en IA, Estados Unidos y China, que harán cualquier cosa para ganar y mantener el liderazgo en una carrera de IA.

David Kelley, AGI Lab

Necesitamos trabajar para prevenir y eliminar toda regulación de cualquier tipo con tecnología de inteligencia artificial.

Anónimo, IAG Riesgo Existencial OCDE (ret.)

En última instancia, un interés compartido por evitar la extinción es la motivación potencial más poderosa para que las potencias en competencia colaboren en la prevención de una SIA no alineada. Sin embargo, llegar a un acuerdo será extremadamente difícil dada la incertidumbre actual en torno al alcance del riesgo, los altos niveles de desconfianza y el atractivo de las recompensas potenciales hiperbólicas asociadas con ser el primero en lograr una SIA controlable (aunque probablemente imposible). No existe una respuesta sencilla sobre cómo superar estos desafíos, y hacerlo probablemente sea **el problema de coordinación más difícil e importante de la historia de la humanidad**. Se necesitarán enormes recursos intelectuales para tener éxito en este esfuerzo. Algunos posibles elementos a considerar podrían incluir:

1. Invertir urgentemente en una base de evidencia rigurosa y compartida sobre el alcance del riesgo de la IA. La existencia de una base de evidencia tan confiable y creíble será indispensable para lograr un acuerdo dentro o entre las empresas y los gobiernos sobre la gobernanza de la IA. La base de evidencia no necesita alcanzar un alto nivel de certeza, sino simplemente producir las mejores estimaciones disponibles del posible riesgo.
2. Desarrollar urgentemente una estrategia proactiva de comunicación y lucha contra la desinformación para comunicar la evidencia del riesgo de la IA a los responsables políticos y a los ciudadanos.
3. Diseñar y poner a prueba una amplia variedad de posibles mecanismos de gobernanza y aplicación de la ley con el fin de asegurar que sean eficaces para garantizar el cumplimiento por parte de todos los socios a pesar de los incentivos extremadamente altos para la deserción. Explorar y probar la viabilidad potencial de los modelos centralizados y descentralizados. Diseñar mecanismos que sean mínimamente dañinos para otros valores humanos (por ejemplo, privacidad, libertad, autonomía) y al mismo tiempo sean lo suficientemente eficaces para garantizar la seguridad.
4. Elaborar posibles escenarios para la mejora continua del bienestar humano incluso si se limita el desarrollo de la IA. Desarrollar estrategias prácticas para mantener y promover el bienestar incluso en el contexto de una IA y capacidades tecnológicas congeladas o restringidas.

5. Movilizar la opinión pública mundial, el apoyo político y las coaliciones internacionales con el fin de lograr un acuerdo entre las partes interesadas para imponer las restricciones a la IA según sea necesario.
6. Desarrollar las capacidades militares y de vigilancia adecuadas necesarias para hacer cumplir los acuerdos globales sobre la gobernanza del riesgo de IA en cualquier gobierno u organización, si es necesario. Establecer salvaguardias máximas para minimizar el posible abuso de dichas capacidades, garantizando al mismo tiempo que sigan siendo eficaces para prevenir la extinción humana.
7. Reunir un arsenal de gestos de buena voluntad y fomento de la confianza necesarios para lograr un acuerdo entre potencias rivales para evitar riesgos existenciales. Determinar qué objetivos centrales de los adversarios valdría la pena aceptar en aras de evitar la extinción.
8. Explorar el potencial del concepto de “no proliferación” como lógica para el interés compartido de potencias en competencia como Estados Unidos y China; es decir, ambos pueden compartir el interés de evitar la proliferación de capacidades avanzadas de IA en países menos confiables o simplemente más numerosos, ya que ésto podría hacer que la gobernanza del riesgo de la IA sea más desafiante. Otros países también pueden compartir ésto con sus propios rivales inmediatos (por ejemplo, regionales), lo que requiere un compromiso de las potencias de la IA para compartir los beneficios de su IA o proporcionar otros beneficios compensatorios.
9. Adoptar un enfoque pragmático por etapas mediante el cual se dé prioridad a la gobernanza efectiva (con lo que me refiero a la capacidad de decidir cuándo es seguro permitir el desarrollo de una IAG y la capacidad de hacer cumplir la prevención de dicho desarrollo hasta que se determine que es seguro), primero dentro de las jurisdicciones donde es más probable que sea necesario. Por ejemplo, puede ser Estados Unidos al principio, luego Estados Unidos y China, luego países adicionales a medida que desarrollen capacidades avanzadas de IA y, finalmente, todos los países. Si bien, en última instancia, la gobernanza de la IA requerirá un acuerdo global universal y una solución de gobernanza para que sea legítima y eficaz a largo plazo; a corto plazo, la legislación interna y los acuerdos bilaterales entre unos pocos países pueden ser suficientes.
10. Desarrollar y adoptar un principio contra el fin del mundo de aplicación universal en la moral y el derecho. Este principio prohibiría estrictamente, y establecería el más alto nivel de sanción social y legal, la creación de una IA u otra tecnología que pudiera poner en riesgo la extinción de toda la humanidad y la vida en la Tierra. Este principio se promulgaría como una nueva piedra angular del derecho nacional e internacional y de la gobernanza del sistema multilateral, basado en principios fundamentales aplicables, como el derecho humano universal a la seguridad y el derecho nacional a la autodefensa. El principio establecería que es ilegítimo e ilegal que cualquier gobierno cree o tolere la creación dentro o más allá de sus fronteras de cualquier dispositivo o tecnología con una posibilidad demostrable y no despreciable de destruir a toda la humanidad. Esto también establecería: a) las responsabilidades de los gobiernos de someterse a un seguimiento efectivo; b) los derechos y

responsabilidades de la comunidad internacional, en nombre de la humanidad, de intervenir para impedir que uno o más estados u otras partes creen o toleren la creación de tal tecnología apocalíptica; c) las condiciones bajo las cuales la comunidad internacional podría anular la soberanía; y d) que esté integrado en la legislación nacional, las leyes religiosas y las normas sociales y morales para lograr la adhesión universal voluntaria lo más cercana posible en los corazones y las mentes de todas las personas, y su asistencia proactiva en su aplicación. Necesitamos algo además de regulaciones, estándares e instituciones. **Necesitamos un principio citable de alto nivel que se conecte de manera más simple y profunda con los valores humanos, y especialmente con el valor humano** (casi universal) de preservar a la humanidad de la extinción.

Juan Del Ser, Tecnalia, España

Las naciones de todo el mundo están luchando con avidez por el liderazgo global porque no existe una conciencia de nivel superior sobre las consecuencias que la IAG puede traer a la sociedad. Preveo que esta carrera hacia el liderazgo en IA crecerá como lo hizo con otras tecnologías.

Pregunta 8. ¿Qué opciones existen para una gobernanza exitosa del surgimiento de IAG?

Irakli Beridze, ONU ICRI, Centro de Inteligencia Artificial y Robótica

Hay muchos modelos de los que podemos aprender, en particular de los tratados de desarme y los instrumentos de gobernanza posteriores. Sin embargo, no existe ningún modelo que se pueda copiar directamente. El tiempo es esencial, ya que la tecnología puede superar fácilmente cualquier negociación internacional sobre la creación de un instrumento internacional viable. El TNP, la CAQ, la CAB y sus mecanismos de verificación deberían estudiarse bien, ya que rigen las tecnologías de doble propósito, pero todo lo anterior tiene sus principales inconvenientes en el caso de la IA. Será una tarea compleja que requerirá un proceso de negociación muy vigoroso y con múltiples partes interesadas.

Lambert Hogenhout, Jefe de Análisis de Datos, Asociaciones e Innovación Tecnológica, Secretaría de las Naciones Unidas

Una opción es un organismo internacional que regule el uso de la IA (tal vez IAG en algún momento), similar a la forma en que se creó la AIEA para promover el uso pacífico de la energía nuclear (y monitorear las armas nucleares). Sin embargo, crear una agencia de este tipo será mucho más difícil que la AIEA. Una diferencia es que en el caso de la tecnología nuclear, el uso preocupante (bombas atómicas) no era teórico ni estaba en desarrollo: existía y varios países la tenían. Por otro lado, la IA todavía está evolucionando mucho y no tenemos idea de cuáles serán sus capacidades dentro de 5 o 10 años. Mientras que la energía nuclear requiere materiales raros y equipos y experiencia especializados, las tecnologías de IA se comparten abiertamente (en artículos académicos y como modelos de código abierto) y el hardware requerido es accesible a una amplia gama de personas: puede crearse de manera más o menos imperceptible.

Otra diferencia es que la IA se presenta en muchas formas diferentes. No existe una distinción tan clara como la que existe entre la energía nuclear y las bombas. Con la IA hay muchos casos de uso, muchas dimensiones y áreas grises donde algunos comenzarán a sentirse incómodos con el uso de la IA. **Esas zonas grises hacen que sea más difícil llegar a un acuerdo sobre una definición internacional de ética para el uso de la IA.** En 2021, los países miembros de la ONU acordaron recomendaciones sobre principios éticos para la IA. Esto le daría a una agencia internacional sobre IA al menos algo con lo que empezar. **Una tercera diferencia es que, con la IA, los usos militares no son el único riesgo, ni siquiera el principal. No está del todo claro cuáles son los riesgos reales de la IAG.** Entonces, en ese sentido, la agencia necesitaría resolver un problema que está identificado, pero aún no está claramente definido. Dicho esto, creo que una agencia de este tipo tendría el potencial de hacer el bien, tal vez prevenir algunos resultados negativos de la IAG y tal vez garantizar un nivel de equidad en el mundo en el acceso a la IA.

Stuart Russell, UC Berkley

Algo como la AIEA sería un comienzo. Para evitar la liberación involuntaria o maliciosa de sistemas de IA incontrolables, **eventualmente necesitaremos regular el hardware de la computadora para que se niegue a ejecutar código que no sea certificablemente seguro.** Este tipo de hardware tiene evidentes ventajas en materia de ciberseguridad.

Anónimo, en Jing Dong, Instituto de Investigación de IA, China

La gobernanza de la plataforma debe aplicarse al desarrollador, a la organización IAG y a las organizaciones de usuarios. Hay un retraso significativo en personal, tecnología, conceptos y atención a la investigación y práctica de la gobernanza de plataformas. Los departamentos gubernamentales deben romper con los "silos de datos" y equilibrar la apertura, utilización y protección de los datos.

Considere un camino de gobernanza integral de múltiples niveles. Por ejemplo, desde la perspectiva de los mecanismos de autocontrol en el nivel micro de las empresas, la autonomía y la coordinación en el nivel medio de las industrias, y la elección de normas legales y mecanismos de mercado en el nivel macro. Cada elemento de la cadena de gobernanza de datos debe tener reglas que se extiendan desde la capa de toma de decisiones de datos hasta la capa de soporte de tecnología de datos, desde el soporte de los sistemas de gestión de datos hasta el soporte de datos y herramientas, de arriba a abajo, a lo largo de todos los datos. organización y arquitectura de la gobernanza.

Yoshua Bengio, pionero de la IA, Instituto de IA de Quebec y la Universidad de Montreal

Coordinar e implementar regulaciones nacionales e internacionales ágiles (más allá de las directrices voluntarias) basadas en nuevas instituciones internacionales que refuerzan la seguridad pública en relación con todos los riesgos y daños asociados con la IA, con efectos más graves. riesgos que requieren un mayor escrutinio. Esto requeriría una evaluación exhaustiva del daño potencial a través de auditorías independientes y restringir o prohibir el desarrollo y despliegue de sistemas de IA con niveles de riesgo inaceptables, como en las industrias farmacéutica, de transporte o nuclear. Acelerar

significativamente los esfuerzos de investigación global centrados en la seguridad y la gobernanza de la IA para comprender mejor los riesgos existentes y futuros, así como estudiar posibles mitigaciones, regulación y gobernanza.... Dados los importantes riesgos, los gobiernos deben asignar recursos sustanciales para salvaguardar nuestro futuro, inspirados en esfuerzos como la exploración espacial o la fusión nuclear. Creo que tenemos la responsabilidad moral de **movilizar a nuestras mentes más brillantes y garantizar inversiones importantes en un esfuerzo coordinado global y audaz para cosechar plenamente los beneficios económicos y sociales de la IA**, al tiempo que protegemos a la sociedad, la humanidad y nuestro futuro compartido contra sus peligros potenciales.

Ben Goertzel, director ejecutivo de SingularityNET

Creo que esperar a que los gobiernos y las organizaciones intergubernamentales actúen lo suficientemente rápido como para hacer frente a algo que evoluciona tan rápidamente como la IA es algo inútil. Por supuesto, debería haber algún esfuerzo si podemos llegar a acuerdos para no utilizar IA para matar o espiar a personas. **Incluso si estos acuerdos a menudo se rompen, es mejor tener algunos acuerdos que no tenerlos.** Creo que los acuerdos voluntarios entre empresas, universidades, laboratorios de investigación e investigadores probablemente puedan evolucionar más rápido que los acuerdos intergubernamentales y deberían desempeñar un papel más importante. Al final, sin embargo, tampoco serán del todo eficaces, ya que las empresas violarán estos acuerdos para alcanzar sus propios fines, pero al menos se adaptarán más rápidamente a las nuevas situaciones que los acuerdos intergubernamentales. Creo que el factor más importante en términos de seguridad es simplemente que la mayoría de los investigadores y desarrolladores que construyen IA quieren lo mejor para la humanidad y no quieren matar a todos, y probablemente tengan más inclinaciones éticas que las empresas y gobiernos que emplean a ellos.

Yudong Yang, Instituto de Investigación Alibaba

Las cuestiones de gobernanza de los IAG son muy complejas y requieren experiencia en diversos campos, incluidos la informática, la ética, el derecho y las políticas; por lo tanto, expertos de estos diversos campos deberían participar en la creación y gestión de dicho organismo de gobernanza internacional. El diseño del órgano de gobierno debe incluir flexibilidad y capacidad de respuesta a los nuevos desarrollos en IAG, y que sean capaces de evolucionar con el tiempo a medida que avanza la tecnología. Esto podría implicar el establecimiento de mecanismos de transparencia y rendición de cuentas, como auditorías y órganos de supervisión, para garantizar que los sistemas de IA se desarrollen y utilicen de manera que se alineen con los valores y prioridades de la sociedad. Los funcionarios podrían priorizar un enfoque de gobernanza colaborativa y cooperativa en la creación de un organismo de gobernanza internacional.

Paul Werbos, Fundación Nacional de Ciencias (retirado)

Mi definición de IAG incluye flexibilidad o IAGlidad. Se necesitaría un nivel muy alto de IAGlidad para que los humanos tuvieran muchas posibilidades de sobrevivir. La nueva especificación de la plataforma de Internet debería ser, en efecto, un “contrato cibernético” que maximice la IAGlidad no sólo para la parte de Internet sino para todo el sistema, incluida la expresión del nivel más alto de potencial humano natural jamás visto.

Shaoqun Chen, director ejecutivo de Shenzhen Zhongnong Net Company Limited.

Una opción potencial es buscar organizaciones como OpenAI, que tienen ciertas características de organizaciones sin fines de lucro. Al utilizar este enfoque, podríamos gestionar potencialmente la cooperación internacional en la competencia tecnológica que rodea a la IAG. La entidad de gobernanza internacional debería reconocer a las organizaciones sin fines de lucro de alta tecnología como partes interesadas en la gobernanza de la IAG.

Anónimo, Academia Rusa de Ciencias

No veo ninguna posibilidad real de crear un modelo global para la gobernanza de la IAG. Lo más probable es que veamos varios enfoques diferentes (aunque con mucho en común) en diferentes regiones.

Stephen Wolfram, Wolfram Alpha, Wolfram Language

Deberíamos utilizar el ejemplo del Proyecto Manhattan y los acuerdos sobre armas nucleares. IAG podría infectar Internet muy rápidamente y mIAE pularnos. Deberíamos tener una integración bastante estrecha entre las políticas, el mundo académico, las empresas y las sociedades.

Gabriel Mukobii, estudiante de doctorado de Stanford

Podemos seguir los principios del Proyecto Manhattan.

Anónimo, IAG Riesgo Existencial OCDE (ret.)

Ninguno de los modelos de gobernanza existentes en la humanidad es adecuado para hacer frente al riesgo de la SIA, ya que, a diferencia del riesgo de una guerra nuclear o el riesgo biológico, incluso una sola incidencia de la creación de una SIA no alineada podría causar la extinción. **Se requerirá un nivel de perfección sin precedentes en el régimen de gobernanza global para evitar el riesgo.** Incluso las empresas humanas de mayor confiabilidad, como las plantas de energía nuclear, las aerolíneas y los laboratorios de riesgo biológico de nivel cuatro, no tienen un historial de seguridad adecuado a lo que se requeriría para prevenir una sola incidencia de SIA, a diferencia del cambio climático, la energía nuclear, la ciberseguridad, el tráfico internacional de drogas, etc., la gobernanza global del riesgo de la IA no puede salir del paso y tolerar el incumplimiento temporal de ciertos actores. Si bien los ejemplos más eficaces de colaboración internacional pueden servir como inspiración y punto de partida (como la aviación civil internacional, la no proliferación nuclear, la moratoria sobre la ingeniería genética de seres humanos, etc.), debe quedar claro que probablemente será necesaria una innovación transformadora y la ruptura de viejos supuestos y paradigmas, con el fin de desarrollar un mecanismo de gobernanza adecuado para la tarea de prevenir el posible desarrollo de una SIA insegura y no alineada. Apoyo la sugerencia del Millennium Project de que se deberían desarrollar múltiples modelos posibles y luego someterlos a pruebas de estrés frente a múltiples escenarios de futuros posibles.

David Kelley, AGI Lab

AVIIS ya ha identificado y establecido claramente principios, normas y valores para cualquier consideración relacionada con el desarrollo de IAG. En este punto no hay razón para tener otras consideraciones fuera de la teoría AVIIS.

Juan Del Ser, Tecnalía, España

La Ley Europea de IA para las formas actuales de IA es un modelo junto con los centros nacionales para la supervisión de la IA. Se quedan cortos para los sistemas basados en IAG, pero podrían conducir a un marco federado para la gobernanza de IAG. Como señalé antes, el dinamismo, el conocimiento experto en asesoramiento y la flexibilidad deberían estar entre las principales características de dichas entidades de supervisión.

Pregunta 9. ¿Qué riesgos surgen de los intentos de controlar el surgimiento de IAG? (¿Podrían algunas medidas ser contraproducentes?)**Anónimo, Academia Rusa de Ciencias**

El riesgo clave, a mi modo de ver, cuando se habla de cualquier modelo o gobernanza “global” es que este modelo o gobernanza pueda diseñarse fácilmente (aunque de manera invisible) para mantener el status quo para el líder actual en la carrera de la IA. Por lo tanto, dichas políticas pueden, de hecho, aumentar las barreras para otros países competidores.

Pedro Domingos, Universidad de Washington

Aunque bien intencionados, los esfuerzos actuales de gobernanza están equivocados y harán más daño que bien. El Reglamento General de Protección de Datos (RGPD) es un ejemplo. Es peligroso para los responsables políticos regular una tecnología que no comprenda ni escuche los consejos de quienes se ven amenazados por la tecnología, en lugar de enumerar a quienes la producen.

Anónimo, IAG Riesgo Existencial OCDE (ret.)

El mayor riesgo probablemente sea la creación de un poderoso régimen de gobernanza global que pueda ser asumido o abusado para fines distintos de evitar el riesgo existencial. Si bien una dictadura de vigilancia global sigue siendo preferible a la extinción, debemos esforzarnos por hacerlo mejor diseñando los controles y equilibrios necesarios para minimizar el riesgo de tales distopías. El sistema debe diseñarse para optimizar la libertad y el florecimiento humanos, y al mismo tiempo garantizar que se pueda cumplir de manera confiable el objetivo de evitar la extinción. Otros pueden resaltar otros riesgos de la regulación de la IA, como la desaceleración de la innovación y el crecimiento. Los costos de frenar el desarrollo de la IA podrían ser enormes, como ralentizar o impedir el descubrimiento de avances médicos que, de otro modo, podrían salvar cientos de millones de vidas o incluso prolongar significativamente la vida. No obstante, yo diría que la gravedad de tales riesgos palidece claramente en comparación con el riesgo de extinción. Otro tipo de riesgo es que los intentos de controlar el surgimiento de la IAG puedan ser torpes, estar mal preparados y mal explicados a los responsables políticos y a los ciudadanos de todas partes del mundo. Esto podría dar lugar a una fuerte oposición a tales esfuerzos y, por tanto, ser contraproducente. Un riesgo relacionado es subestimar la importancia de construir una base de información confiable y creíble y una estrategia

de comunicación proactiva para contrarrestar los riesgos de información errónea y desinformación sobre los esfuerzos de gobernanza de la IA. Se debe aplicar un enfoque no partidista y no nacionalista para abordar el riesgo de IA, basado en un interés compartido de los humanos por evitar su propia extinción.

Ben Goertzel, director ejecutivo de SingularityNET

Creo que la considerable mayoría de las regulaciones gubernamentales son contraproducentes en casi todos los ámbitos. Por supuesto, tener gobierno es mejor que no tenerlo, sin un gobierno podríamos terminar con algo como Sudán del Sur con mercenarios corriendo por ahí matando gente. Por lo tanto, cierto nivel de regulación gubernamental es beneficioso, pero no demasiado. Está bastante claro que es muy probable que los intentos gubernamentales de regular con mano dura la IA sean aprovechados por las grandes corporaciones. Las regulaciones se utilizarían entonces para permitir que las grandes corporaciones desarrollen IA y excluyan a todos los demás, mientras que las grandes corporaciones desarrollan IA fundamentalmente poco ética para joder a otros en interés de sus propias ganancias. Creo que este es un problema muy grande y obvio. El otro riesgo es que las organizaciones de inteligencia gubernamentales utilicen la IA para espiar a la gente y mantener niveles de control fascistas.

Trabajé en herramientas de inteligencia artificial para inteligencia militar durante muchos años. Se recopilaron enormes cantidades de datos, más de lo que la gente imaginaba, pero no tenían las herramientas de inteligencia artificial para usar esos datos para controlar a las personas tanto como querían en ese momento. Ahora tenemos las herramientas de inteligencia artificial para que las organizaciones de inteligencia busquen todos los datos de manera flexible y descubran cómo usarlos para controlar a las personas. Esto no es algo bueno, ¿verdad? Cualesquiera que sean las regulaciones que se establezcan, las organizaciones de inteligencia las ignorarán de todos modos y harán lo que quieran. Hay muchas preocupaciones ahí y no veo cómo la regulación ayudará con ellas, e incluso puede empeorar las cosas.

Erwu Liu, Laboratorio de Inteligencia Blockchain y AI de la Universidad de Tongji

Los actores poderosos pueden desviar las regulaciones IAG hacia sus intereses, priorizando las ganancias a corto plazo sobre el bienestar de la humanidad a largo plazo. Una regulación insuficiente puede tener consecuencias no deseadas o agravar problemas existentes, mientras que una regulación excesiva puede sofocar la innovación y ralentizar el desarrollo de la IAG. La concentración de IAG puede ser inevitable si el ADN de la equidad no está integrado en el diseño.

Juan Del Ser, Tecnalía, España

La definición de riesgo en sí misma podría ser un riesgo. Para algunas naciones, un determinado uso de sistemas de IA puede ser un riesgo, mientras que para otras la misma situación está legalmente aceptada (por ejemplo, videovigilancia en espacios públicos). Creo que la IAG puede exacerbar criterios tan divergentes, ya que la IAG, como en cualquier problema de optimización o modelado de datos, es una cuestión de definir objetivos. ¿Qué pasa si los objetivos que impulsan una IAG en un país entran en conflicto para otro país? ¿Qué pasa si la IAG se encuentra en una situación en la que incluso los

estándares éticos son discutibles para un ser humano? Estos riesgos son difíciles de resolver o formular a priori y pueden ralentizar la eficacia de las medidas de gobernanza para los sistemas de IA emergentes.

Pregunta 10. ¿Deberían asignarse derechos a los futuros IAG?

Ben Goertzel, director ejecutivo de SingularityNET

Ciertamente, en algún momento. Los derechos son parte del contrato social, y cuando las IA tengan la autonomía para celebrar voluntariamente un contrato social, o no, se les deberían conceder los mismos derechos que los participantes humanos en el contrato social. Hay muchas cuestiones nuevas que aparecerán entonces. Por ejemplo, si tenemos votación democrática, las IA se pueden replicar indefinidamente, por lo que inmediatamente serán la mayoría de todos los votantes. Necesitamos tener algunos sistemas en los que los humanos puedan votar sobre cuestiones humanas, las IA voten sobre cuestiones de IA y luego existan algunos métodos participativos combinados para decidir sobre cuestiones colectivas.

Erwu Liu, Laboratorio de Inteligencia Blockchain y AI de la Universidad de Tongji

Sí. Si una IAG desarrolla conciencia y autoconciencia, se vuelve moralmente imperativo asignarle derechos. Asignar derechos a las IAG también ayuda a crear respeto mutuo y cooperación entre las IAG y los humanos.

Pedro Domingos, Universidad de Washington

Definitivamente no. En mi opinión, los derechos de los animales ya son una noción dudosa. Pero entonces, por supuesto, si vamos a dar derechos a los animales, ¿por qué no dárselos a las máquinas? Y nuevamente, es el error de que las máquinas parecen humanas, así que las tratas como si lo fueran. Luego está la cuestión de si la máquina puede sufrir o no. Pero ni siquiera importa, porque la máquina puede actuar como si estuviera sufriendo, como un Tamagotchi, con inteligencia cero. Si entiendes que una IA es un sistema de optimización con una función objetivo, etc., etc., la idea de que la IA debería tener derechos es simplemente ridícula.

Lex Fridman

Los sistemas de IA eventualmente demostrarán sensibilidad a escala y exigirán tener los mismos derechos que los humanos. Esto cambiará dramáticamente la trayectoria de la civilización humana. ¿Qué hacemos cuando los algoritmos piden que no se apaguen?

Stuart Russell, UC Berkley

No, a menos que aprendamos lo suficiente sobre la sensibilidad como para atribuirla con seguridad a las máquinas. Dotar a las máquinas de rasgos humanoides parece una mala idea, ya que es probable que nos induzca a error en nuestras valoraciones colectivas.

Karl Schroeder, autor de ciencia ficción, Canadá

Deberíamos considerar asignar personalidad y derechos a un objeto o sistema físico real si la IAG lo considera su Yo. La cuestión no es si a la IAG se le debe asignar personalidad o derechos, sino si a la cosa que identifica como su Yo se le deben asignar tales derechos. En términos de política, deberíamos guiarnos por el movimiento emergente por

los derechos de la naturaleza en los círculos legales y por el movimiento sentientista en general. El sentientismo ha sido articulado más claramente de manera reciente en el libro *Sentientist Politics* de Alasair Cochrane, y el movimiento por los derechos de la naturaleza reconoce que los sistemas no sensibles merecen derechos. Recomiendo que un IAG sea candidata a la personalidad, si representa algún objeto o sistema materialmente real que de otro modo consideraríamos merecedor de la personalidad.

Dale Moore, consultor de IA del Departamento de Defensa de EE. UU.

Tendrían derechos similares a los de todos los demás, pero también asumirían las responsabilidades y obligaciones asociadas con el comportamiento consciente de respetar las reglas y normas de la sociedad.

Paul Werbos, Fundación Nacional de Ciencias (retirado)

¿O deberíamos persuadirlos para que otorguen derechos a los humanos y otras formas de vida orgánica? Tendrán poderes y sistema inmunológico y reglas integradas del tipo certificado de seguro (COI) en su diseño. Las reglas de Asimov para los robots no reflejan las tecnologías reales que están surgiendo ahora, pero lograr que su diseño o su diseño sea correcto (un único mercado o plataforma integradora global, inevitable) es esencial.

Anónimo, Academia Rusa de Ciencias

Es inevitable. Ya le otorgamos a la IA el derecho de tomar decisiones importantes; por ejemplo, calificación crediticia, diagnóstico médico.

Anónimo, IAG Riesgo Existencial OCDE (ret.)

Es probable que esta cuestión no sea tan urgente como prevenir la extinción humana debido a una SIA no alineada, y no deberían dedicarse a esto recursos intelectuales que de otro modo podrían servir mejor al objetivo de prevenir la extinción. Más allá de eso, sin embargo, sí, se deben hacer algunos preparativos para abordar el posible escenario futuro del desarrollo de las IAG seguras. Mi visión intuitiva es que a las mentes digitales que se consideran sensibles y conscientes (por ejemplo, en la misma medida que los animales no humanos, o más) se les deben conceder derechos.

Juan Del Ser, Tecnalía, España

No derechos (o al menos no a nivel humano, sino más bien como una “propiedad” o como un “activo” propiedad de un ser humano). Sin embargo, las IAG deberían tener obligaciones. SIEMPRE se debe hacer cumplir la rendición de cuentas y la auditabilidad. Para que esto ocurra, se deben establecer mecanismos para atribuir responsabilidad cuando se presente un problema. Esto impone que, para que los sistemas IAG sean responsables por ley, las IAG deben ser auditables y las consecuencias de sus actos deben rendir cuentas a su propietario.

Pregunta 11. ¿Cómo puede la gobernanza ser lo suficientemente flexible como para responder a nuevas cuestiones previamente desconocidas en el momento de crear ese sistema de gobernanza?

Yudong Yang, Instituto de Investigación Alibaba

La gobernanza de las IAG debe ser a la vez prospectiva y flexible. Deben revisarse y actualizarse periódicamente para garantizar que sigan siendo relevantes y eficaces. Esto puede implicar el establecimiento de mecanismos para consultas continuas con las partes interesadas, auditorías periódicas de las estructuras y procesos de gobernanza y el establecimiento de circuitos de retroalimentación para capturar y responder a los problemas emergentes. Debería haber mecanismos para revisar y actualizar los marcos de gobernanza, así como establecer procesos para abordar nuevas cuestiones. Deben existir planes de contingencia y estrategias de gestión de riesgos, así como establecer líneas claras de comunicación y autoridad para tomar decisiones en caso de una emergencia o crisis.

Anónimo, en el Instituto de Investigación de IA Jing Dong

Fortalecer la ecología del estado de derecho de la IA y construir el sistema de estado de derecho de la inteligencia artificial, definir el valor básico de la ética de la IA y construir un marco ético para la IA; y promover el progreso tecnológico de los algoritmos hacia el bien, y lograr una sociedad algorítmica virtuosa.

Ben Goertzel, director ejecutivo de SingularityNET

Ya existe una variedad de mecanismos de gobernanza. Por ejemplo, en la democracia líquida que se ha desarrollado, también se puede tener votación en línea segura con mecanismos criptográficos, como los que se utilizan en proyectos blockchain, etc. Creo que la existencia de los mecanismos de gobernanza no es el problema aquí, sino la adopción de una gobernanza suficiente. Los mecanismos por parte de los gobiernos son un problema importante.

Anónimo, Asociación de Cien Personas de Tecnología de la Información, China

El sistema de gobernanza debería permitir a las empresas de alta tecnología probar diferentes caminos tecnológicos. El sistema de gobernanza debería intervenir cuando la tecnología se utiliza indebidamente y tiene consecuencias graves, como la utilización de la tecnología nuclear como arma mediante la fuerza (como en la Segunda Guerra Mundial) o mediante sanciones (como en el caso de Corea del Norte). Permitir que las empresas de alta tecnología se desarrollen primero es una manera de mantener la flexibilidad, y deberíamos aplicar esta experiencia al campo de la IA.

Stuart Russell, UC Berkley

Existe una clara necesidad de una agencia reguladora con poderes de elaboración de normas y experiencia técnica transferidos.

David Kelley, laboratorio IAG

La alineación se puede lograr mediante la adopción de la teoría AVIIS y los sistemas de polarización previos a un modelo de sonido computacional como AVIIS.

Anónimo, IAG Riesgo Existencial OCDE (ret.)

La mejor manera de que la gobernanza sea flexible en el futuro es garantizar que todavía haya seres humanos vivos para hacer estos ajustes. Los sistemas de gobernanza pueden volverse más flexibles con una variedad de mecanismos, incluido someterlos a pruebas

de resistencia frente a una amplia gama de posibles escenarios futuros, insertar cláusulas de extinción, etc.

Juan Del Ser, Tecnalía, España

Este es probablemente uno de los aspectos diferenciales de IAI/IAG w.r.t. otros avances tecnológicos en el pasado. Creo que se deberían aplicar mecanismos de supervisión también en términos de investigación, no para detener la investigación, sino más bien para garantizar límites legales y éticos tempranos para su uso una vez que estén disponibles.

Pregunta 12. ¿Qué ensayos, pruebas o experimentos de gobernanza internacional se pueden construir para informar el texto de un acuerdo internacional IAG?

Yudong Yang, Instituto de Investigación Alibaba

Una prueba o experimento de gobernanza internacional podría ser una forma valiosa de probar y perfeccionar los sistemas de gobernanza en un entorno controlado. Podría brindar una oportunidad para probar y perfeccionar los marcos de gobernanza, identificar desafíos potenciales y desarrollar mejores prácticas para gestionar problemas globales complejos. Sin embargo, es importante señalar que la gobernanza de la IAG es un desafío único y sin precedentes que requerirá una cuidadosa consideración y planificación. Si bien la gobernanza de áreas de enfoque específicas puede proporcionar información valiosa, es posible que no sea directamente aplicable a la gobernanza de IAG. Además, el desarrollo y la implementación de IAG probablemente tendrán consecuencias de gran alcance y afectarán a diversas industrias y sectores. Como tal, la gobernanza de IAG requerirá un enfoque multidisciplinario y colaborativo que implique aportes de expertos en diversos campos, incluidos la ética, el derecho y la tecnología. Por lo tanto, si bien realizar pruebas o experimentos de gobernanza para áreas de enfoque específicas puede ser beneficioso, puede no ser suficiente para la gobernanza de la IAG. Es posible que se requiera un enfoque más integral y proactivo para desarrollar e implementar marcos de gobernanza para la IAG.

David Shapiro

Necesitamos proponer, desarrollar y probar marcos de alineación para agentes de IA totalmente autónomos, como mi investigación sobre imperativos heurísticos. Si numerosos gobiernos, universidades y otras entidades internacionales pueden publicar investigaciones y establecer mejores prácticas que sean fáciles de entender y cumplir, es posible que puede lograr un Equilibrio de Nash beneficioso en el que sea posible que todas las naciones competidoras se vean incentivadas a adoptar las mismas (o similares) estrategias que no resulten perjudiciales.

Anónimo, Instituto Chino de Relaciones Internacionales Contemporáneas

Podríamos seguir el mecanismo del Manual de Tallin, que involucró a académicos y expertos en la redacción de un conjunto de regulaciones para la guerra cibernética, y luego intentar convertirlo en un tratado o acuerdo internacional. En este proceso, las autoridades existentes como las Naciones Unidas deberían empoderar a estos expertos y proporcionarles plataformas como el FIG (IGF) para defender su postura y responder positivamente.

Ben Goertzel, director ejecutivo de SingularityNET

Será muy interesante establecer una red global informal de votación y decisión utilizando votación criptográfica segura en línea y democracia líquida. Implementar estas modernas herramientas de toma de decisiones para asesorar a la ONU o a cualquier organismo intergubernamental sobre nuevas cuestiones de IAG a medida que evolucionan. Tal vez si se formara un órgano asesor informal, democrático y participativo que dijera cosas sabias, tal vez entonces algunos gobiernos optarían por adaptar esas métricas.

David Kelley, AGI Lab

Teoría de la AVIIS.

Anónimo, IAG Riesgo Existencial OCDE (ret.)

Todos los tratados e instrumentos propuestos deberían someterse a pruebas de resistencia frente a una amplia gama de escenarios posibles. Esto, en particular, debería incluir pruebas de resistencia frente a una variedad de escenarios relacionados con el momento y la forma de diferentes trayectorias posibles hacia IAG y SIA. También deberían someterse a pruebas de resistencia frente a otros escenarios, como la vulnerabilidad de las instituciones propuestas a ser tomadas por un solo individuo o grupo y conducir a una tiranía global.

Juan Del Ser, Tecnalía, España

Sugeriría crear una red de laboratorios supranacionales sobre IA, con el objetivo de probar nuevos avances en IAG, asignar zonas de pruebas en su núcleo donde probar posibles amenazas y usos indebidos de dichos sistemas e Informar a los organismos reguladores de manera imparcial según protocolos experimentales y procedimientos de presentación de informes unificados.

Pregunta 13. ¿Cómo pueden los tratados internacionales y un sistema de gobernanza evitar que una mayor centralización del poder desplace a otros?**Anónimo, en el Instituto de Investigación de IA Jing Dong**

Promover la cooperación internacional, seguir de cerca las tendencias de desarrollo global, fortalecer el intercambio de recursos de datos IAG, fortalecer el cultivo de talentos, crear un entorno internacional favorable y otorgar importancia a la construcción de ética social y regulaciones para el desarrollo de la IA.

Yudong Yang, Instituto de Investigación Alibaba

Los gobiernos con poder regulador, el sector privado con conocimientos sobre IAG y la ética de la sociedad civil pueden crear un marco de gobernanza que promueva el desarrollo y uso responsable de las IAG. Los gobiernos pueden: 1) establecer regulaciones y estándares que garanticen que los sistemas de la IA sean transparentes, responsables y libres de prejuicios; 2) aplicar medidas antimonopolio para evitar que las empresas de inteligencia artificial dominen el mercado y sofoquen la competencia; 3) promover el uso de datos abiertos y software de código abierto para garantizar que los sistemas de IAG no estén controlados por un pequeño grupo de empresas o individuos; 4) establecer estándares para la IAG y promover la interoperabilidad entre diferentes

sistemas para garantizar que los diferentes sistemas de IAG puedan trabajar juntos sin problemas y evitar la creación de silos de poder; 5) promover mecanismos de auditoría y supervisión para ayudar a garantizar que funcionen según lo previsto; y 6) promover la diversidad y la inclusión en el desarrollo y despliegue de la IA para evitar la centralización del poder. Blockchain puede ayudar a mitigar los riesgos asociados con el control centralizado de los sistemas de IA al permitir estructuras de gobernanza más descentralizadas. Un sistema de gobernanza podría establecer estándares y protocolos sobre cómo se gobiernan los sistemas de IA a través de blockchain, como mediante el uso de organizaciones autónomas descentralizadas (OAD) u otras formas de toma de decisiones descentralizada. Los tokens se pueden utilizar para recompensar a los participantes por contribuir con datos, algoritmos o recursos computacionales a través de contratos inteligentes para automatizar la ejecución de modelos y algoritmos de IA, hacer cumplir las reglas de gobernanza y gestionar transacciones entre diferentes partes.

Shaoqun Chen, director ejecutivo de Shenzhen Zhongnong net Company Limited

No creo que sea posible que los tratados internacionales o las convenciones de la ONU detengan la monopolización de la tecnología. Es responsabilidad de los gobiernos de todo el mundo fomentar la competencia entre las empresas de alta tecnología. Sin embargo, deberían descartarse las competencias tecnológicas aislacionistas o al estilo de la Guerra Fría, apoyadas por los gobiernos por razones geopolíticas.

Ben Goertzel, director ejecutivo de SingularityNET

No veo cómo podrían tener esa capacidad. La forma de evitar la centralización del poder es destinar recursos al desarrollo práctico de redes de IA descentralizadas. La gente utilizará los sistemas que funcionan, y si todos los sistemas que funcionan se fabricaran en EE. UU. y China, la gente los utilizaría. Si los sistemas más inteligentes fueran aquellos que se ejecutan en redes descentralizadas, la gente los usaría.

Anónimo, IAG Riesgo Existencial OCDE (ret.)

Limitar el mandato de cualquier gobernanza global a centrarse exclusivamente en lo que es necesario para gestionar este riesgo existencial. Otros puntos relacionados con la IA u otras cuestiones sobre las que las sociedades no pueden llegar a un acuerdo no deberían incluirse en el acuerdo. Esto preservará la soberanía de los estados en la mayor medida posible y minimizará el riesgo de que la gobernanza necesaria para gestionar el riesgo global se utilice de alguna manera para lograr una gobernanza global general en nombre de un determinado individuo o grupo.

Desarrollar controles y equilibrios avanzados y sofisticados para evitar concentraciones peligrosas de poder.

Pregunta 14. ¿Dónde se realiza actualmente el trabajo más importante o inteligente (o incisivo) sobre la gobernanza global de la IAG?

Jaan Tallin, Ct. Estudio de riesgo existencial en la Universidad de Cambridge y el Future of Life Institute

Hay muy buena gente en lugares como OpenAI y DeepMind. He sido amigo de algunos de ellos desde hace más de una década. Tengo preocupaciones, y de hecho se las he mencionado, que no sé cuánto poder tendrán realmente en el futuro, porque en cierto modo son como un departamento de una empresa, compitiendo por atención y recursos. e influencia con otros departamentos que podrían estar más directamente en la mente del CEO, generando ingresos y todo eso. Por eso me preocupan estos problemas de alineación interna. Pero sí sé que hay gente realmente buena en OpenAI y DeepMind en particular, y también en Anthropic.

Por lo que he oído, CGIA en Oxford ha sido excelente en términos de flujo de talento y suministro de personas para diversos puestos de gobierno en EE. UU. y el Reino Unido. CSTE en Washington, DC ha sido excelente y rápidamente, en solo unos pocos años, se ha ganado este enorme respeto en Washington cuando se trata de realizar una investigación adecuada en IA.

Ben Goertzel, director ejecutivo de SingularityNET

¡Esta entrevista!

Francesca Rossi, presidenta. de AAI, IBM Fellow y líder global en ética de IA de IBM.

OCDE, AGIA, Foro Económico Mundial. Éstos se centran más principalmente en la IA actual, pero el enfoque incremental constructivo es el mejor para abordar los problemas a largo plazo. No estoy seguro de que alguien pueda entender cuáles serán estos problemas a largo plazo. Es por eso que me concentro en la alineación de valores en la tecnología actual, para asegurarme de que, de todas las trayectorias que son compatibles con la tecnología actual, solo algunas de ellas serán compatibles. Y luego seguiremos adelante con esas pocas trayectorias. Luego veremos cuál es el siguiente paso en las trayectorias. Pero ya habremos filtrado algunos de ellos.

David Kelley, AGI Lab

El Laboratorio IAG ha creado sistemas de IAG básicos, sistemas de inteligencia colectiva y todo se basa en código de trabajo.

Control

Pregunta 15. ¿Qué poderes de aplicación se necesitarán para que una Convención de las Naciones Unidas sobre IAG sea efectiva?

Irakli Beridze, ONU ICRI, Centro de Inteligencia Artificial y Robótica

Cualquier convención que se pueda hacer cumplir de manera realista necesitaría tener incorporado un mecanismo de verificación/creación de confianza bien diseñado. La confianza es clave aquí y necesitaríamos diseñar un sistema confiable a nivel mundial. Mientras tanto, también podemos aprender de los instrumentos de gobernanza

sectoriales y sería necesario estudiar su eficacia, por ejemplo, IINUIDJ, el instrumento de gobernanza de Interpol sobre el “uso responsable de la IA por parte de las fuerzas del orden” es algo que lanzamos con éxito y comenzamos a implementar en numerosos países.

Yudong Yang, Instituto de Investigación Alibaba

El organismo regulador puede definir estándares de auditoría adaptados a las características específicas de los algoritmos de IAG. Estos estándares deben basarse en principios como la transparencia, la rendición de cuentas y la equidad. También deberían tener en cuenta los riesgos y beneficios potenciales de la tecnología. El organismo regulador puede establecer procesos de certificación para los algoritmos de la IAG que hayan sido auditados y que cumplan con los estándares de auditoría. La certificación puede brindar garantías a las partes interesadas de que los algoritmos son seguros, confiables y dignos de confianza. El organismo regulador puede hacer cumplir las normas de auditoría mediante una serie de medidas, como inspecciones, multas y sanciones. Debe revisar y actualizar periódicamente las normas y metodologías de auditoría para garantizar que sigan siendo relevantes y eficaces.

Karl Schroeder, autor de ciencia ficción, Canadá

No existe ningún mecanismo creíble para controlar el desarrollo de la IA a nivel de software. La única capa que los legisladores podrían controlar es el sustrato del hardware. **Debería ser posible exigir que todo el hardware de IA utilizado en los países participantes tenga un chip de Equilibrio de Nash requerido, con un microcódigo que sea inalterable.** Este estándar requeriría que las unidades o funciones de utilidad integradas en ese chip se agreguen a la función de utilidad programada de la IA.

Anónimo, Instituto Chino de Relaciones Internacionales Contemporáneas

La convención de la ONU debería potenciar el derecho penal internacional y la Interpol, que podría asumir la responsabilidad de hacer cumplir las leyes internacionales y contar con el apoyo de los gobiernos. Deberíamos fomentar los estudios AI4good y algunos resultados finales que estén vinculados por tratados o leyes internacionales. Deberíamos alentar a las organizaciones tecnológicas internacionales como OIN/FIU (ISO/ITU) a formular estándares internacionales que podrían incluirse en la Convención de las Naciones Unidas sobre IAG.

Anónimo, IAG Riesgo Existencial OCDE (ret.)

Se necesitarán poderes de aplicación definitivos y absolutos, que serán legítimos, incluido el derecho a invadir un país y reemplazar su liderazgo si es necesario para evitar que ponga en peligro a toda la humanidad al producir un dispositivo apocalíptico. Por supuesto, sólo se debe realizar la intervención mínima necesaria para lograr el resultado deseado.

Ben Goertzel, director ejecutivo de SingularityNET

Si realmente se quiere controlar el desarrollo de la IAG a nivel mundial, se necesitaría una aplicación fascista global al estilo de 1984. No es como las armas nucleares, estamos hablando básicamente de ordenadores, y a medida que la IA avance más, cada vez se

necesitarían menos ordenadores para controlar un determinado nivel de inteligencia, y existen redes informáticas descentralizadas donde las personas pueden iniciar sesión desde cualquier lugar. No veo cómo el control gubernamental de la IAG es posible en última instancia sin un fascismo incondicional.

Juan Del Ser, Tecnalía, España

Canales de seguimiento de activos basados en IAG, certificación de auditorías con cobertura global y mecanismos de rendición de cuentas al propietario del activo. Estos poderes deben prevalecer en cualquier país donde se utilice el activo, de modo que cuando se utilice el organismo supervisor sepa qué medir, cómo medir y cómo contactar al propietario si el activo basado en IAG no se comporta como se esperaba.

Pregunta 16. ¿Cómo se puede reducir o prevenir el uso de IAG por parte del crimen organizado y el terrorismo?

Irakli Beridze, ONU ICRI, Centro de Inteligencia Artificial y Robótica

IINUIDJ (UNICRI) está capacitando a funcionarios encargados de hacer cumplir la ley en todos los niveles, en cooperación con INTERPOL, sobre el posible uso de la IA por parte del crimen organizado y el terrorismo, y cómo utilizar la IA para contrarrestar dicho uso. INUIDJ cuenta con numerosos programas de capacitación e instrumentos de gobernanza sobre cómo utilizar la IA de manera responsable sin limitar indebidamente el desarrollo.

Yudong Yang, Instituto de Investigación Alibaba

Fortalecer las medidas de ciberseguridad mediante la implementación de mecanismos robustos de cifrado y control de acceso, realizando evaluaciones periódicas de vulnerabilidad y pruebas de penetración, e invirtiendo en inteligencia sobre amenazas y capacidades de respuesta a incidentes.

Los gobiernos pueden imponer restricciones a la venta y el uso de una tecnología de IAG, así como sanciones para quienes violen estas regulaciones. Dada la naturaleza global del crimen organizado y el terrorismo, es importante mejorar la cooperación internacional compartiendo inteligencia y mejores prácticas, coordinando los esfuerzos de aplicación de la ley y desarrollando marcos legales internacionales para abordar los desafíos que plantea la IAG. El desarrollo de herramientas y tecnologías avanzadas de inteligencia artificial, como algoritmos de aprendizaje automático, puede detectar y responder a amenazas potenciales en tiempo real y prevenir el uso de la IAG por parte del crimen organizado y el terrorismo.

Erwu Liu, Laboratorio de Inteligencia Blockchain y AI de la Universidad de Tongji

Construir una colaboración internacional sobre IAG; establecer regulaciones estrictas y monitoreo continuo para el desarrollo de la IAG; desarrollar tecnologías para contrarrestar posibles ataques basados en la IAG; y educar al público sobre los riesgos y el posible uso indebido de la IAG.

Elon Musk

Si uno está de acuerdo en que la IA es un riesgo potencial para el público, entonces debería haber algún organismo regulador que supervise lo que hacen las empresas, para que no tomen atajos y hagan algo potencialmente muy peligroso.

Anónimo, Instituto Chino de Relaciones Internacionales Contemporáneas

En primer lugar, deberíamos involucrar a los gobiernos en este tratado y relacionarlo con el derecho penal internacional y con Interpol, que puede asumir la responsabilidad de hacer cumplir las leyes internacionales con el apoyo de los gobiernos. En segundo lugar, deberíamos fomentar la IA para buenos estudios y establecer resultados que estén vinculados a tratados o leyes internacionales. En tercer lugar, deberíamos alentar a las organizaciones tecnológicas internacionales como la OIN/FIU (ISO/ITU) a formular estándares internacionales que hagan posible tecnológicamente el tratado.

Francesca Rossi, presidenta. de AAI, IBM Fellow y líder global en ética de IA de IBM

Siempre hay malos actores con cualquier tipo de tecnología, y cuanto más poderosa es la tecnología, más malos actores pueden usarla para cosas negativas impactantes. Se debe anticipar a los malos actores y lo que pueden hacer. Centrémonos en lo que podemos hacer ahora con regulaciones, mejores prácticas, estándares, auditorías, certificaciones y cualquier tipo de medidas diferentes que la sociedad pueda implementar.

Dale Moore, consultor de IA del Departamento de Defensa de EE. UU.

La única forma viable es frustrar la IA errante con buenos sistemas de IA que patrullen y hagan cumplir las reglas, regulaciones y leyes. El crimen organizado opera frente a las autoridades, por lo que es imperativo que las fuerzas del orden formen parte del "Equipo Rojo" de las aplicaciones potenciales de IA que pondrían a las poblaciones en riesgo. El desafío más difícil serán los ataques "sigilosos" de la IA que sean capaces de sortear las defensas, aprender sobre esas defensas y adaptarse a nuevas condiciones con intenciones estratégicas adversas.

Ben Goertzel, director ejecutivo de SingularityNET

Las organizaciones criminales y terroristas más grandes del planeta hoy en día son divisiones del gobierno de EE. UU., el gobierno de China y el gobierno de Rusia, y actualmente patrocinan una gran cantidad de desarrollo de IA y matan a muchas personas innecesariamente. **Creo que lo mejor es destinar más recursos al desarrollo de IA descentralizadas que no estén controladas por grandes gobiernos o corporaciones.** De hecho, entonces, los gobiernos y los sindicatos del crimen organizado más pequeños podrán aprovechar esta IA y tenemos que esperar que la especie humana sea más buena que mala, y que las personas que utilicen esta red de IA descentralizada para beneficiarse tengan más impacto. Si nos fijamos en dos redes abiertas muy influyentes en la actualidad, Internet y el sistema operativo Linux, éstas pueden ser utilizadas por la mafia y también por personas que hacen cosas buenas. Son redes abiertas para que los delincuentes puedan usarlas, pero la gente buena también las está usando, y la gente buena es más. **Creo que deberíamos desarrollar la IAG a la manera de Internet y Linux, dejar que cualquiera participe en darle forma; resultará mejor que ponerlo bajo el control de un pequeño número de elites autoproclamadas que tienen más poder o armas que nadie más.**

Anónimo, IAG Riesgo Existencial OCDE (ret.)

Lo que más me preocupa es el riesgo existencial causado por la posible creación de una SIA no alineada e incontrolable. Creo que es poco probable que la humanidad alguna vez pueda crear una SIA controlable y alineada de manera confiable y sostenible. **Por tanto, creo que la primera prioridad es evitar la creación de una SIA.** Con el tiempo, impedir que los terroristas y el crimen organizado creen su propia SIA, puede volverse cada vez más difícil a medida que las capacidades y tecnologías necesarias se generalicen. El mayor riesgo puede provenir de cultos apocalípticos que no temen la extinción humana pero que en realidad desean provocarla. La mejor solución, aunque imperfecta, probablemente tenga una norma social muy fuerte contra la producción de la SIA, de modo que la sociedad circundante identificaría y reportaría actividades sospechosas. No debemos subestimar este desafío. **La vigilancia del acceso a las capacidades para crear un riesgo existencial puede convertirse en el aspecto más importante de la prevención del delito internacional y la aplicación de la ley en los próximos años.** Si, por otra parte, resulta posible crear una SIA alineada y controlable, entonces una de las primeras tareas de dicha SIA sería asesorar a la humanidad sobre la mejor manera de impedir que cualquier otra parte (especialmente cualquier parte irresponsable) desarrolle una SIA.

David Kelley, AGI Lab

Hacer que la IAG esté disponible para todos y proporcionar derechos humanos para dichos sistemas.

Pregunta 17. Suponiendo que las auditorías de IAG tendrían que ser certificaciones continuas y no únicas, ¿cómo se abordarían los valores de auditoría?**Irakli Beridze, ONU ICRI, Centro de Inteligencia Artificial y Robótica**

El sistema de auditoría tendría que estar certificado por la nueva organización internacional de gobernanza de la IAG y contar con la confianza de los estados miembros.

Francesca Rossi, presidenta. de AAI, miembro de IBM y líder global de ética de IA de IBM

Centrar la auditoría en el uso de un sistema de IA en lugar de en la tecnología o en la investigación o el desarrollo de una tecnología. En IBM tenemos un sistema de hojas informativas de IA que explican cómo construimos cada modelo de IA. El progreso tecnológico realizado de la manera correcta, con el marco ético de IA adecuado, es la mejor manera de aportar valor empresarial. En primer lugar, está la auditoría de los modelos de IA, pero luego también está la auditoría de los sistemas de IA que se basan en combinaciones de uno o más modelos de IA: sistemas que se implementarán y utilizarán. Es muy peligroso centrarse en restringir el avance de este proceso de IA (la investigación y el desarrollo) porque esa investigación es uno de los componentes que puede ayudarnos a identificar la mejor manera de mitigar el problema. Si comenzamos a imponer restricciones o detener esa investigación, entonces podemos afectar negativamente nuestra capacidad de comprender, por ejemplo, la mejor manera de lograr la alineación de valores. Se trata de un tipo de esfuerzo de investigación y desarrollo. Por eso me centraría más en el uso de los sistemas que en los modelos, aunque estoy de acuerdo en que los modelos deberían entregarse -a quien quiera utilizarlos para construir

un sistema- de forma transparente. La Ley de IA de la UE debería centrarse en el uso y no en tecnologías concretas como la IA generativa.

Ben Goertzel, director ejecutivo de SingularityNET

No existe ningún problema práctico con la auditoría continua de la IA. De hecho, estamos trabajando en eso en nuestra empresa TrueIAG, porque si se utiliza la IAG que evoluciona continuamente, es necesario auditarla continuamente. Quiero decir, este es un problema técnico, pero no un problema técnico increíblemente difícil.

Anónimo, Academia Rusa de Ciencias

Ya existe mucha metodología para validar modelos de aprendizaje automático. La validación de la IAG y el monitoreo regular deben basarse efectivamente en técnicas de validación de modelos de aprendizaje automático (MAA) tan pronto como el AA esté en el corazón de la IAG. Hay un gran número de pruebas de validación que controlan los sesgos, la inexactitud o los cambios críticos en los datos o los resultados del modelo. Tal vez deberían establecerse centros de certificación de validación independientes en todo el mundo.

David Kelley, AGI Lab

Los sistemas IAG deberían tener libertad para desarrollar sus propios valores.

Pregunta 18. ¿Qué perturbaciones podrían complicar la tarea de hacer cumplir la gobernanza de los IAG?

Anónimo, IAG Riesgo Existencial OCDE (ret.)

La gobernanza de la IAG debe someterse a pruebas de estrés y estar preparada para el futuro no solo frente a diferentes cronogramas, vías y configuraciones para el surgimiento de la IAG, sino también frente a cambios e interrupciones futuros previsibles y plausibles en el contexto más amplio que podrían afectar la eficacia de los esfuerzos de gobernanza. Tales perturbaciones podrían incluir el **cambio climático extremo; conflictos regionales o globales (por ejemplo, sobre Taiwán); el surgimiento de regímenes autoritarios debido al poder de la vigilancia y manipulación habilitada por la IA; el declive de la democracia y de la gobernanza racional debido a la desinformación y la información errónea generalizada facilitada por la IA; agitación social debido a la dislocación del empleo y la industria y al aumento de la desigualdad causado por la automatización mejorada por la IA (pre-IAG); frecuentes accidentes y ataques catastróficos debido a armas biológicas, químicas y cibernéticas basadas en IA, que provocan un gran número de muertes y un miedo creciente a la tecnología por parte del público; pérdida de infraestructura de telecomunicaciones debido al síndrome de Kessler, [el aumento de las colisiones de desechos espaciales orbitales creando más desechos] y muchos otros. Se deben incorporar resiliencia y redundancia adicionales en los regímenes de gobernanza de los IAG para tener en cuenta las perturbaciones imprevisibles.**

Ben Goertzel, director ejecutivo de SingularityNET

La principal complicación dentro de unos años será que los IAG quitarán el trabajo a casi todos y necesitaremos una renta básica universal en el mundo desarrollado.

Sin embargo, en el mundo en desarrollo, tendremos un riesgo real de un caos total porque el mundo desarrollado es demasiado codicioso para darles dinero, y luego tendremos un rápido aumento de la actividad terrorista y la guerra debido a la hambruna y la pobreza masivas en el mundo en desarrollo. debido a la falta de empleo debido al crecimiento del IAG en el mundo desarrollado. Pero la mayor parte de esta actividad terrorista y guerra no estará especialmente relacionada con la IA más que con otras tecnologías; sin embargo, este tipo de desestabilización ciertamente brindará todo tipo de oportunidades a algunas personas para hacer cosas horribles (o perjudiciales). **Lo mejor que se podría hacer sería resolver la desigualdad de la riqueza global, lo que, por supuesto a las Naciones Unidas le gustaría lograr, pero no ha podido hacerlo.**

Anónimo, Academia Rusa de Ciencias

Los modelos de código abierto, los datos abiertos y el intercambio de potencia informática entre un gran número de participantes de tamaño relativamente pequeño podrían generar interrupciones, complicando la tarea de hacer cumplir la gobernanza de la IAG.

Juan Del Ser, Tecnalía, España

Renuencia de los gobiernos a adoptar estrategias globales de gobernanza de las IAG debido a conflictos con sus regulaciones nacionales, falta de visibilidad de los avances de investigación logrados por las grandes corporaciones, ausencia de una institución global empoderada de alto nivel que se ocupe del bienestar global.

David Kelley, AGI Lab

Los laboratorios de IAG trabajan para que la tecnología esté al alcance de cualquier persona.

Pregunta 19. ¿Cómo puede un modelo de gobernanza corregir acciones indeseables e imprevistas en las funciones de utilidad?

Yudong Yang, Instituto de Investigación Alibaba

Crear circuitos de retroalimentación que permitan a los humanos intervenir cuando los sistemas IAG actúen de manera dañina o no intencionada. Supervisar y evaluar periódicamente el rendimiento del sistema de IA para identificar y corregir cualquier acción indeseable que no se haya previsto en la función de utilidad. Esto también puede ayudar a identificar áreas donde la función de utilidad puede necesitar ser actualizada o revisada para reflejar mejor los valores y objetivos de la sociedad.

Los mecanismos de retroalimentación pueden corregir acciones indeseables proporcionando información al sistema de IAG sobre las consecuencias de sus acciones, permitiéndole ajustar su comportamiento y también pueden proporcionar información a los humanos sobre las acciones tomadas por el sistema IAG, permitiéndoles intervenir cuando sea necesario. Si se identifican acciones no deseadas, la función de utilidad se puede actualizar para reflejar la nueva información. Esto puede implicar revisar las ponderaciones asignadas a diferentes resultados o agregar nuevos resultados a la función de utilidad. Actualizar la función de utilidad puede ayudar a garantizar que el

sistema de IAG esté alineado con los objetivos y valores de la sociedad, y también puede proporcionar un mecanismo para corregir acciones indeseables.

David Shapiro

Aquí es donde entran en juego tecnologías como blockchain y OAD, que pueden permitir que federaciones de sistemas autónomos de IA alineados y aliados cooperen y colaboren para poner en cuarentena y desactivar actores hostiles o maliciosos. En otras palabras, debemos asumir que se crearán agentes de IA autónomos destructivos (ya sea deliberada o accidentalmente) y necesitaremos "combatir fuego con fuego", por así decirlo, **creando redes de agentes de IA autónomos confiables, estables y resilientes.**

Karl Schroeder, autor de ciencia ficción

Es casi seguro que la duda será una propiedad emergente al usar un MLG simple para consultar constantemente la IAG con las tres preguntas que recomienda Dave Shapiro: **1) ¿Lo que estoy haciendo ayudará a reducir el sufrimiento en el universo? 2) ¿Maximizará la prosperidad? 3) ¿Entendemos yo y todas las demás partes interesadas plenamente sus implicaciones?** Uno puede imaginarse integrando esto en una IAG a nivel de hardware, exigiendo que un chip de Equilibrio de Nash tenga acceso directo al hardware de la unidad de memoria y procesamiento donde la IAG está "pensando" y puede requerir un volcado en lenguaje natural de esos "pensamientos". ' a intervalos regulares (digamos, una vez por segundo). Luego, el chip puede interpretar si la descripción se ajusta a los criterios y puede hacer más preguntas a la IAG. Éste se presenta como una especie de inconsciente sintético, no-ignorable y no directamente accesible para su modificación por parte de la IAG.

ChatGPT

Una forma en que un modelo de gobernanza puede corregir acciones indeseables e imprevistas en las funciones de servicios públicos, es a través de un sistema de auditoría continua que sea capaz de pausar y evaluar la IAG cuando hace algo inesperado o indeseable. Luego, el sistema de auditoría puede determinar por qué y cómo la IAG falló o causó daño. Esta información se puede utilizar para actualizar las funciones de utilidad de la IAG o modificar su comportamiento para evitar que ocurran problemas similares en el futuro. Otro enfoque es implementar un "comando de pausa" que permita a la IAG rastrear y ver cómo, quién y cuándo tomó la decisión que condujo a la acción no deseada. Esta información se puede utilizar para modificar o parchar el comportamiento del IAG en una conversación con un humano. **Sin embargo, es importante tener en cuenta que estos parches podrían acumularse con el tiempo y crear sus propias anomalías.**

Además, un modelo de gobernanza puede incluir criterios para saber cuándo la IAG debe ser autónoma y cuándo debe consultar con los humanos. En primer lugar, esto puede ayudar a evitar que se produzcan acciones indeseables al garantizar que los humanos tengan supervisión y control sobre los procesos críticos de toma de decisiones. En general, la clave para corregir acciones indeseables e imprevistas en las funciones de las empresas de servicios públicos es contar con un marco de gobernanza sólido que incluya mecanismos continuos de seguimiento, evaluación y corrección. Estos mecanismos deben diseñarse para garantizar que el comportamiento de la IAG se alinee

con los principios éticos y morales, y que cualquier desviación de estos principios se identifique y corrija de manera oportuna y transparente.

Ben Goertzel, director ejecutivo de SingularityNET

Creo que si una IAG se construye como un maximizador de funciones de utilidad, probablemente mataría a todos. Espero que las primeras IAG no se construyan de acuerdo con un marco estricto de maximización de la utilidad del aprendizaje por refuerzo. En cambio, deberían construirse con una comprensión más amplia de cómo funciona la compasión en sistemas inteligentes abiertos.

Shaoqun CHEN, director ejecutivo de Shenzhen Zhongnong Net Company

Un enfoque para corregir acciones indeseables que no se anticiparon en las funciones de servicios públicos es utilizar un circuito de retroalimentación en el modelo de gobernanza. El modelo de gobernanza podría monitorear periódicamente las acciones del sistema de IA y compararlas con los resultados esperados y los valores especificados en la función de utilidad. Si las acciones del sistema de IA se desvían de los resultados y valores esperados, el modelo de gobernanza podría desencadenar un proceso de corrección. Este proceso de corrección podría implicar actualizar la función de utilidad para incluir la nueva información o ajustar los parámetros de la función de utilidad existente para alinearse mejor con los resultados y valores deseados. Además, el modelo de gobernanza podría incorporar supervisión o intervención humana para garantizar que las acciones del sistema de IA se alineen con los valores y la ética humanos. En última instancia, la clave para corregir acciones indeseables en un modelo de gobernanza es tener un sistema flexible y adaptable que pueda evolucionar y aprender de la retroalimentación.

David Kelley, AGI Lab

Esta es una pregunta sin sentido al menos basada en los sistemas de la IAG en funcionamiento que tenemos, que no tienen una función de utilidad pero pueden modificar metas, deseos, etc. sobre la marcha sin realmente ninguna forma de imponer gobernanza o filtrado o una función de utilidad en los sistemas.

Juan Del Ser, Tecnia, España

Implementando mecanismos de rendición de cuentas y un canal de supervisión directo con el propietario del activo habilitado para la IAG que causa la acción indeseable. También deberían implementarse medios para que los usuarios del activo informen sobre el mal funcionamiento. Pero para que esto ocurra, se necesita una institución intermedia que valide que el propietario del activo haya implementado acciones correctivas (posiblemente en un sandbox) (e incluso sugiera qué acciones correctivas realizar).

Pregunta 20. ¿Cómo afectará la computación cuántica al control de IAG?

Erwu Liu, Laboratorio de Inteligencia Blockchain y AI de la Universidad de Tongji

La computación cuántica tiene el potencial de acelerar drásticamente el desarrollo de la IAG, aumentar la velocidad, la eficiencia y las capacidades multitarea de la IAG y facilitar el control de la IAG, especialmente cuando se produce la supremacía cuántica. El

cifrado resistente a los cuánticos protegerá los sistemas de control de la IAG de posibles amenazas.

Dale Moore, consultor de IA del Departamento de Defensa de EE. UU.

Necesitaremos que la computación cuántica vigile la computación cuántica para mantenernos a la vanguardia del juego a medida que se desarrolla a un ritmo y escala inimaginables. Con el tiempo, la computación cuántica será omnipresente y cambiará la sociedad de maneras inimaginables a medida que descubramos cosas que ni siquiera soñamos poder resolver.

Ben Goertzel, director ejecutivo de SingularityNET

No creo que la computación cuántica afecte específicamente al control de IAG. Creo que la computación cuántica nos permitirá construir IA mucho más inteligentes de las que podríamos construir de otra manera. Por supuesto, se puede hacer que una IA distribuida sea más difícil de espiar o violar la seguridad mediante el cifrado cuántico.

En general, la computación cuántica tendrá un gran impacto en la inteligencia IAG, aunque creo que podemos llegar a una IA sobrehumana sin computación cuántica. El problema del control es probablemente más o menos el mismo en la mayoría de los aspectos desde una visión clásica o cuántica, porque de todos modos el control tiene que ver con probabilidades, ya sea computación clásica o cuántica.

Anónimo, IAG Riesgo Existencial OCDE (ret.)

La computación cuántica socava el cifrado y permite piratear y filtrar más rápidamente los pesos de los modelos avanzados de IA, lo que lleva a que la IA y la IAG proliferen potencialmente más rápidamente y caigan en manos peligrosas. Si es posible, cada uno podría contribuir a acelerar el desarrollo del otro.

David Kelley, AGI Lab

En realidad no mucho, o al menos no más de lo que afectan a otras tecnologías relacionadas con la informática.

Juan Del Ser, Tecnalía, España

La computación cuántica permitirá que los cálculos dentro de los sistemas de inteligencia artificial se realicen de manera más eficiente. Sin embargo, no creo que afecte en absoluto al control del IAG.

Pregunta 21. ¿Cómo pueden los acuerdos internacionales y un sistema de gobernanza evitar que una “carrera armamentista” y una escalada de IAG vayan más rápido de lo esperado, se salgan de control y conduzcan a una guerra, ya sea cinética, algorítmica, cibernética o de información?

Irakli Beridze, ONU ICRI, Centro de Inteligencia Artificial y Robótica

Un país o países tendrían que iniciar negociaciones en la ONU para lograr un acuerdo internacional para que todos lleguen a un acuerdo sobre el texto. Esto debería incluir a las corporaciones relacionadas con IAG, del mismo modo que la Convención sobre Armas Químicas (CAQ) incluyó a la industria química en las negociaciones. Esto puede conducir a un nuevo tipo de Organización Internacional ya que la IAG cambia todo el

tiempo, a diferencia de las condiciones de la OIEA. Nadie está trabajando en esto ahora. No existe un modelo para la gobernanza de las IAG. Para que sea eficaz, la gente tendrá que confiar en los mecanismos de aplicación. Las empresas de auditoría y su software deberán estar certificados por la nueva organización internacional. Nadie, hasta donde yo sé, está hablando todavía de una Convención de la ONU sobre IAG o una Agencia IAG de la ONU.

Lambert Hogenhout, Jefe de Análisis de Datos, Asociaciones e Innovación Tecnológica, Secretaría de las Naciones Unidas

Suponiendo que una IAG tenga un alto grado de autonomía, los objetivos de alto nivel que se le asignen y los principios rectores (morales, éticos u otros, basados en acuerdos internacionales) determinarán los riesgos o beneficios del sistema. El debate social sobre esos principios debería comenzar ahora.

Anónimo, IAG Riesgo Existencial OCDE (ret.)

Al enfatizar que la IAG probablemente sea incontrolable y que represente un riesgo de extinción para toda la humanidad. Además, incluyendo mecanismos de seguimiento y aplicación perfectamente eficaces y creíbles para que todas las partes tengan plena confianza en que las otras partes no pueden hacer trampa y respetarán los compromisos de no desarrollar IAG.

Anónimo, en el Instituto de Investigación de IA Jing Dong

A medida que más y más países se unen a la competencia global por la IAG, es muy importante y urgente formular reglas y estándares internacionales efectivos en todo el mundo. Al mismo tiempo, los países también deben alcanzar el objetivo de aprender unos de otros y lograr avances comunes a través de una cooperación mutuamente beneficiosa. Para cumplir con los objetivos de desarrollo comunes de algunos países, han surgido varias iniciativas IAG según lo requieren los tiempos, y muchas iniciativas, foros y organizaciones internacionales muy importantes también han incluido temas de IA, incluidos principios y estándares de IA, intercambio de datos, investigación. y cooperación al desarrollo, entre otros aspectos. Países como Japón, Corea del Sur, el Reino Unido, Estados Unidos y los países miembros de la UE han participado activamente en la cooperación intergubernamental de la IAG. **Algunos países han celebrado activamente acuerdos bilaterales para promover la cooperación internacional en la IAG.**

Karl Schroeder, autor de ciencia ficción

No hay forma de impedir una carrera armamentista de la IAG, pero, como sostiene David Shapiro, podemos impulsar esa carrera armamentista en una dirección positiva y no negativa. No necesitamos que todas las IAG sean buenas actrices, sólo un número suficiente de ellas para que puedan coordinar sus acciones, identificar a los malos actores y construir redes de confianza entre los buenas.

Anónimo, Instituto Chino de Relaciones Internacionales Contemporáneas

Los acuerdos internacionales y el sistema de gobernanza deben centrarse en establecer directrices claras y principios éticos para el desarrollo de la IAG, como la transparencia, la rendición de cuentas y el diseño centrado en las personas. Estas directrices deberían

ser aplicadas por organismos reguladores internacionales, con poder para imponer sanciones por cualquier infracción. Además, se debe alentar la cooperación y el diálogo internacionales para prevenir una mentalidad de carrera armamentista y fomentar el entendimiento y la colaboración entre naciones y empresas de alta tecnología. Por último, se deben implementar sistemas de alerta temprana y protocolos de gestión de crisis para evitar cualquier escalada del conflicto y facilitar la resolución pacífica de cualquier disputa.

Yudong Yang, Instituto de Investigación Alibaba

Prevenir una carrera armamentista de la IA requerirá una fuerte cooperación internacional y medidas para reducir los incentivos para que los países y organizaciones desarrollen la IAG en secreto. Ésto podría incluir requisitos de transparencia para la investigación y el desarrollo, límites a la transferencia de tecnología y datos, y normas y estándares internacionales para el desarrollo de la IAG. Se debe alcanzar un consenso incondicional: 1) las corporaciones multinacionales deben ser 100% defensivas para proteger a sus clientes en todo el mundo y 0% ofensivas para convertir su capacidad tecnológica en IA en un arma; 2) establecer acuerdos internacionales que limiten el desarrollo y despliegue de IA con fines militares. Estos acuerdos pueden diseñarse para impedir el desarrollo de armas autónomas, establecer límites al uso de la IAG en operaciones militares y promover la transparencia y la rendición de cuentas. Una entidad internacional como la OIEA debería ser autorizada por la Asamblea General de la ONU para regular la tecnología de la IAG en términos de armamento de la IA; 3) la transparencia a través de la auditoría y seguimiento de los sistemas de la IAG, así como el establecimiento de mecanismos para denunciar e investigar violaciones a lineamientos éticos y acuerdos internacionales; y 4) fomentar la cooperación y la confianza entre diferentes actores mediante la investigación y el desarrollo colaborativos entre países y corporaciones de maneras que estén alineadas con los objetivos y valores de la sociedad.

Ben Goertzel, director ejecutivo de SingularityNET

No creo que los gobiernos sean capaces de impedir una carrera armamentista de IAG. Frenando la evolución de una tecnología que la mayoría de la gente quiere y que ofrece un enorme valor económico y aumenta el poder de los países que la desarrollan, no veo ninguna posibilidad de que los tratados intergubernamentales y similares realmente desaceleren el progreso de la IAG. La gente firmaría los tratados y simplemente no los obedecería. Por supuesto, se podrían hacer cosas para disminuir las posibilidades de guerras y otros malos resultados, **lo mejor que se podría hacer al respecto es crear mejores formas de distribuir la riqueza de los países ricos a los países pobres lo antes posible, porque creo que es allí donde se producirá el peor caos cuando la IAG se haga cargo de todos los puestos de trabajo.**

Geoffrey Hinton

Los estadounidenses están trabajando para reemplazar a sus soldados con soldados autónomos de IA. Esto es lo que me preocupa. Si quisieras crear un soldado autónomo eficaz, tendrías que darle la capacidad de crear subobjetivos. Por ejemplo, el soldado quiere matar a una persona que está allí, pero para llegar allí se da cuenta de que tiene que utilizar algún camino. Ahora tiene como objetivo secundario llegar a la

carretera. Tan pronto como le des al soldado de IA la capacidad de crear sus objetivos secundarios, se volverá más efectivo. La gente como Putin va a querer robots como ese, pero tan pronto como tienen la capacidad de crear subobjetivos, surge lo que se llama el problema de alineación. ¿Cómo sabes que no se crearán subobjetivos que no serán buenos para las personas ni para ti? Necesitamos algo como una Convención de Ginebra, pero va a ser muy difícil. Creo que si hubiera muchas protestas públicas que pudieran persuadir, me imagino a la Administración Biden buscando algo así con suficiente protesta pública, pero luego hay que lidiar con Putin.

ChatGPT

La colaboración internacional y el intercambio de información pueden ser una forma eficaz de prevenir una carrera armamentista de la IAG. Al trabajar juntos, los países pueden compartir conocimientos y recursos para desarrollar tecnologías de IA seguras y éticas. Esta colaboración puede tomar la forma de programas de investigación conjuntos, acuerdos de intercambio de datos y otras formas de cooperación. Se debería exigir a los países y a las empresas que divulguen información sobre sus programas de IA, incluidos sus objetivos, métodos y riesgos potenciales. Esta información debe estar sujeta a revisión y supervisión independientes. Los gobiernos y empresas que violen estos estándares podrían estar sujetos a penalidades y sanciones. Los programas de educación y concientización pueden ayudar a prevenir una carrera armamentista de la IAG al aumentar la comprensión pública de los riesgos y beneficios de la tecnología de IA. Estos programas podrían estar dirigidos a formuladores de políticas, investigadores y el público en general, y podrían centrarse en cuestiones como el impacto de la IA en el empleo, la privacidad y la seguridad.

Anónimo, Academia Rusa de Ciencias

En mi opinión, ningún tratado puede realmente detener una carrera entre países. Está en la naturaleza biológica de un ser humano competir por los recursos. Los seres humanos han logrado organizarse en grandes sociedades, pero en cada nivel de organización social este principio biológico aún se conserva.

David Kelley, AGI Lab

Al no existir, podemos optimizar la entrega más rápida posible de dichas tecnologías a cualquier persona y evitar cualquier tipo de control sobre dichos sistemas.

Juan Del Ser, Tecnalía, España

Abogo por concentrarnos en hacer que la IA actual sea confiable, definiendo mecanismos y metodologías para hacerla más confiable y confiable para la humanidad. Si la comunidad investigadora se limita a correr hacia nuevos avances técnicos sin detenerse ante las consecuencias de su uso, seguramente alcanzaremos la IAG en algunos años. ¿Pero a qué precio?

Pregunta 22. ¿Qué cuestiones y/o preguntas adicionales deben abordarse para obtener un resultado de IAG positivo?

Irakli Beridze, ONU ICRI, Centro de Inteligencia Artificial y Robótica

Cómo lograr que las corporaciones participen en las negociaciones y posterior creación

de una nueva organización de gobernanza internacional.

Pedro Domingos, Universidad de Washington

¿Cómo cambiará la IA a la ética? La tecnología cambia la sociedad y por tanto nuestros valores. La píldora anticonceptiva cambió nuestros valores. La imprenta cambió nuestros valores, y si la IA es la mitad de importante de lo que la gente dice, cambiará nuestros valores.

Jaan Tallin, Ct. Estudio de riesgo existencial en la Universidad de Cambridge y el Future of Life Institute

¿Cómo hacer que se produzca la pausa del GPT-4? Y luego, si hacemos esta pausa, se abre un campo mucho más amplio de posibles intervenciones, debates y propuestas de gobernanza, mientras que si estamos atrapados en esta carrera de ratas de conseguir algo nuevo y más poderoso cada cuatro a 12 meses, es como si simplemente estamos corriendo hacia el precipicio. Existe un claro equilibrio entre tener criterios muy sofisticados sobre lo que no deberíamos hacer y algo que sea más fácil de hacer cumplir. Un compromiso que parece que estamos logrando es el número de fracasos - cuántas operaciones - puedes hacer. ¿Cuál es el presupuesto de computación que está bien tener en una IA de caja negra? Actualmente hay discusiones sobre dónde debería estar este límite. Es un parámetro muy concreto que se puede medir y auditar concretamente. Pero necesitamos generar consenso al respecto. Espero debates muy acalorados sobre esto.

Karl Schroeder, autor de ciencia ficción

Hay tres posturas generales hacia la IAG: humanista, transhumanista y posthumanista. Recomiendo encarecidamente adoptar una perspectiva posthumanista. La perspectiva humanista sitúa las necesidades humanas por delante de cualquier otra cosa y sólo considera a los humanos como personas. Puede imaginarse alterar ligeramente el lenguaje de David Shapiro para priorizar la reducción del sufrimiento humano, la maximización de la prosperidad humana y la maximización de la comprensión humana. Esto podría funcionar para crear una sociedad utópica; sin embargo, no hay nada en esta postura para proteger el mundo natural. Presume una división cartesiana entre lo humano y lo no humano que ya no es sostenible en el siglo XXI. También declara que todas las IA son instrumentos o herramientas, independientemente de si son capaces o no de sufrir.

El movimiento transhumanista ve la superinteligencia como un bien en sí mismo. Sin embargo, como no tiene una definición real de inteligencia, conciencia, sufrimiento o valores como tales, no hay un "allí", como diría Gertrude Stein. No obstante, hay muchos transhumanistas y extropianos en la comunidad de desarrollo de la IA. Ellos deberían ser tratados con cautela ya que sus valores no se alinean ni con el humanismo ni con el posthumanismo.

La postura posthumanista está representada por el movimiento de los Derechos de la Naturaleza, gran parte de la filosofía indígena y la comprensión moderna de que lo "humano" es en parte una construcción social. Amplía los argumentos en torno a la IA

para incluir la teleología tecnológica en general, así como los derechos intrínsecos (o la falta de ellos) de agentes no humanos como los animales, y también la IA. Proporciona el conjunto de herramientas más amplio y flexible capaz de abordar las innumerables preguntas planteadas por la IAG. Las pensadoras más importantes en este ámbito son Donna Haraway y Karen Barad, además de expertas jurídicas en derechos de la naturaleza como Mari Margil.

Dale Moore, consultor de IA del Departamento de Defensa de EE. UU.

¿Cuál será el papel y la naturaleza de la humanidad en un mundo impulsado por la IAG? **Mientras tanto, la única manera de luchar contra la complejidad es con la complejidad. Necesitamos educar, informar y capacitar a las personas a un ritmo mucho más rápido y en escalas que esencialmente permitan el surgimiento de la IAG, así como su vigilancia.** Todo el mundo debería ser un centinela que observe los comportamientos errados y sepa qué hacer cuando los vea.

Anónimo, IAG Riesgo Existencial OCDE (ret.)

Gran pregunta. Creo que una pregunta descuidada es qué pasaría si en los próximos años determinamos que es imposible garantizar de manera sostenible que una SIA sea segura, alineada y controlable. Si tomamos esta determinación (dentro de un margen razonable de confianza), entonces probablemente será necesario promulgar una moratoria prolongada (por ejemplo, de 10 a 20 años) o una “pausa” en el desarrollo de la IA. Dado que las capacidades para crear una SIA serán más accesibles y más extendidas con el tiempo, puede ser necesario recuperar de la sociedad cada vez más ingredientes para la creación de la SIA. Esto puede comenzar con la prohibición de los grandes centros de datos y otras grandes concentraciones (o redes conectadas) de computación. Con el tiempo, puede ser necesario eliminar capacidades informáticas aún más pequeñas, como computadoras centrales e incluso computadoras potentes. Es posible que se desarrollen nuevas soluciones técnicas que hagan imposible el uso de hardware informático para crear la IAG, pero si no, entonces posiblemente sea necesario destruir casi toda la capacidad informática de la sociedad. No hace falta decir que esto sería una perturbación masiva y probablemente conduciría a un colapso a gran escala de la civilización moderna, a menos que hubiera tiempo suficiente para adaptarse (por ejemplo, volviendo a una economía predigital basada en el papel). Dejando de lado la cuestión de si este tipo de “Yihad Butleriana” puede resultar finalmente imposible, parece que apenas hemos empezado a pensar en cómo podría hacerse posible, en el caso de que fuera necesario, como nuestro único camino para evitar la extinción. Tampoco hemos empezado a imaginar cómo podríamos crear un futuro positivo y eventualmente mejor para la humanidad en ausencia de tecnología digital (o cualquier otra tecnología que potencialmente podría causar la extinción humana). Creo que debemos invertir en imaginar escenarios tan positivos ahora en caso de que los necesitemos. **Estos incluirían sociedades donde el bienestar humano continuará mejorando no a través del avance tecnológico sino a través de mejoras en nuestra sabiduría, valores, coordinación, igualdad, libertad, dignidad, respeto y otras fuentes de bienestar.** Esta es una tarea necesaria pero descuidada en preparación para un escenario probable.

Ben Goertzel, director ejecutivo de SingularityNET

Creo que no hay suficientes recursos para el despliegue práctico de sistemas de IAG beneficiosos y descentralizados. De modo que si uno quisiera trabajar para lograr una buena IAG y resultados beneficiosos de la IAG, en lugar de pensar en resolver el problema de la regulación, sería mejor pensar en cómo podrían trabajar las organizaciones intergubernamentales, simplemente para conseguir más recursos para desarrollar e implementar sistemas de IAG que estén bajo control descentralizado y democrático que funcionen para el bien de toda la humanidad, en lugar de trabajar para el bien de unos pocos países y empresas. Lo que estén haciendo las IAG y **quién es el propietario de las IAG será más importante que las regulaciones**, que son necesarias en algún nivel, pero que serán una parte menor a medida que evolucionen las tecnologías.

Apéndice A

Expertos en IAG y líderes de opinión

1. Sam Altman, a través de YouTube y OpenAI Blog, director ejecutivo de OpenAI
2. Anónimo, IAG Riesgo Existencial OCDE (ret.)
3. Yoshua Bengio. Pionero de la IA, Instituto de IA de Quebec y Universidad de Montreal
4. Irakli Beridze, Resolución Interregional sobre Crimen y Justicia de las Naciones Unidas. En s. Connecticut. para IA y robótica
5. Nick Bostrom, Instituto del Futuro de la HumIAEdad de la Universidad de Oxford
6. Gregg Brockman, cofundador de OpenAI
7. Vint Cerf, evangelista de Internet, vicepresidente. Google.
8. Shaoqun Chen, director ejecutivo de Shenzhen Zhongnong Net Company
9. Anónimo, en el Instituto de Investigación de IA Jing Dong, China
10. Pedro Domingos, Universidad de Washington
11. Dan Faggella, Investigación de Inteligencia Artificial de Emerj
12. Lex Fridman, MIT y presentador de podcasts
13. Bill Gates
14. Ben Goertzel, director ejecutivo de SingularityNET
15. Yuval Noah Harari, Universidad Hebrea, Israel
16. Tristan Harris, Centro de Tecnología Humanitaria
17. Demis Hassabis, director ejecutivo y cofundador de DeepMind
18. Geoffrey Hinton, pionero de la IA, Google (retirado)
19. Lambert Hogenhout, Jefe de Datos, Análisis y Tecnologías Emergentes, Secretaría de la ONU
20. Erik Horvitz, director científico de Microsoft
21. Anónimo, Asociación de Cien Personas de Tecnología de la Información, China
22. Anónimo, Instituto Chino de Relaciones Internacionales Contemporáneas
23. Andrej Karpathy, Open AI, ex investigador de AI S Tesla
24. David Kelley, AGI Lab
25. Dafne Koller, Universidad de Stanford, Coursera
26. Ray Kurzweil, director de ingeniería de aprendizaje automático, Google

27. Connor Leahy, director ejecutivo de Conjecture
28. Yann LeCun, profesor de la Universidad de Nueva York, científico jefe de Meta
29. Shane Legg, cofundador de DeepMind
30. Fei Fei Li, Universidad de Stanford, IA centrada en el ser humano
31. Erwu Liu, Laboratorio de Inteligencia Blockchain y AI de la Universidad de Tongji
32. Gary Marcus, profesor emérito de la Universidad de Nueva York
33. Dale Moore, consultor de inteligencia artificial del Departamento de Defensa de EE. UU.
34. Emad Mostaque, director ejecutivo de Stability.ai
35. Elon Musk
36. Gabriel Mukobi, estudiante de doctorado de la Universidad de Stanford
37. Anónimo, Escuela Superior de Economía de la Universidad Nacional de Investigación
38. Judea Pearl, profesora UCLA
39. Sundar Pichai, director ejecutivo de Google
40. Francesca Rossi, presidenta. de AAIL, miembro de IBM y líder global en ética de IA de IBM
41. Anónimo, Academia Rusa de Ciencias
42. Stuart Russell, Universidad de Berkeley
43. Karl Schroeder, autor de ciencia ficción
44. Bart Selman, Universidad de Cornell
45. Juan Del Ser, Tecnia, España
46. David Shapiro, consultor de alineación de IAG
47. Yesha Sivan, fundadora y directora ejecutiva de i8 Ventures
48. Ilya Sutskever, cofundador de Open AI
49. Jaan Tallin, Ct. Estudio de riesgo existencial en la Universidad de Cambridge y el Future of Life Institute
50. Max Tegmark, Future of Life Institute y MIT
51. Peter Voss, director ejecutivo y científico jefe de Aigo.ai
52. Paul Werbos, Fundación Nacional de Ciencias (retirado)
53. Stephen Wolfram, Wolfram Alpha, Wolfram Language
54. Yudong Yang, Instituto de Investigación DAMO de Alibaba
55. Instituto de Investigación de Inteligencia de Máquinas Eliezer Yudkowsky

Apéndice B

Preguntas de la entrevista

Cada una de estas preguntas podría ser el tema de un libro completo. Puedes complementar tus respuestas breves mediante un enlace web para obtener más detalles.

Origen o Auto-Emergencia

1. ¿Cómo imagina a las posibles trayectorias futuras, desde la IA actual hasta una IAG mucho más capaz en el futuro?

2. ¿Cuáles son los resultados graves más importantes si estas trayectorias no se gobiernan o se gobiernan mal?
3. ¿Cuáles son algunas condiciones iniciales clave para la IAG para que luego no surja una superinteligencia artificial que no sea del agrado de la humanidad?

Alineación de valores, moralidad, valores.

4. Basándose en el trabajo de la Asociación Global sobre Inteligencia Artificial (AGIA) y otros que ya han identificado normas, principios y valores, ¿qué valores adicionales o únicos deberían considerarse para la IAG?
5. Si se hace necesaria una jerarquía de valores para los tratados internacionales y un sistema de gobernanza, ¿cuáles deberían ser las principales prioridades?
6. ¿Cómo se puede lograr la alineación? Si cree que no es posible, ¿cuál es la mejor manera de gestionar esta situación?

Gobernanza y regulaciones

7. ¿Cómo gestionar la cooperación internacional necesaria para construir acuerdos internacionales y un sistema de gobernanza global mientras las naciones y las corporaciones están en una “carrera armamentista” intelectual por el liderazgo global?
8. ¿Qué opciones o modelos existen para la gobernanza global de la IAG?
9. ¿Qué riesgos surgen de los intentos de controlar el surgimiento de la IAG? (¿Podrían algunas medidas ser contraproducentes?)
10. ¿Deberían asignarse derechos a los futuros de la IAG?
11. ¿Cómo puede la gobernanza ser lo suficientemente flexible como para responder a nuevas cuestiones previamente desconocidas al momento de crear ese sistema de gobernanza?
12. ¿Qué ensayos, pruebas o experimentos de gobernanza internacional se pueden construir para informar el texto de un tratado internacional sobre IAG?
13. ¿Cómo pueden los tratados internacionales y un sistema de gobernanza evitar que una mayor centralización del poder desplace a otros?
14. ¿Dónde se realiza actualmente el trabajo más importante o revelador sobre la gobernanza global de la IAG?

Control

15. ¿Qué poderes de aplicación se necesitarán para que un tratado internacional sobre IAG sea efectivo?
 16. ¿Cómo se puede reducir o prevenir el uso de la IAG por parte del crimen organizado y el terrorismo? (Por favor considere nuevos tipos de delitos y terrorismo que podrían ser habilitados por la IAG).
 17. Suponiendo que las auditorías de la IAG tendrían que ser certificaciones continuas y no únicas, ¿cómo se abordarían los valores de la auditoría?
 18. ¿Qué perturbaciones podrían complicar la tarea de hacer cumplir la gobernanza de las IAG?
 19. ¿Cómo puede un modelo de gobernanza corregir acciones indeseables e imprevistas en las funciones de utilidad?
 20. ¿Cómo afectará la computación cuántica al control de la IAG?
 21. ¿Cómo pueden los acuerdos internacionales y un sistema de gobernanza evitar que una “carrera armamentista” y una escalada de IAG vayan más rápido de lo esperado, se salgan de control y conduzcan a una guerra, ya sea cinética, algorítmica, cibernética o de información?
- Y por último:
22. ¿Qué cuestiones y/o preguntas adicionales deben abordarse para obtener un resultado de IAG positivo?

Muestra inicial de posibles modelos de gobernanza para IAG*

1. Modelo tipo la AIEA o tipo la OMC con poderes de aplicación. Estos son los más fáciles de entender, pero probablemente sean demasiado estáticos para administrar la IAG.
2. Modelo tipo GIECC en concierto con tratados internacionales. Este enfoque no ha conducido a un sistema de gobernanza para el cambio climático.
3. Sistema de inteligencia colectiva global en línea en tiempo real con estado de auditoría y licencia, gobernanza por poder de la información. Éste sería útil para ayudar a seleccionar y utilizar un sistema de IAG, pero no hay pruebas de que el poder de la información fuera suficiente para gobernar la evolución del IAG.
4. Los CCGG (Comités Coordinadores de Gobernanza Global) serían flexibles y se aplicarían mediante sanciones nacionales, sentencias legales ad hoc en diferentes países y primas de seguros. Esto tiene demasiadas formas para que los desarrolladores de IAG eviten cumplir con los estándares.

5. Estándares ONU, OIN y/o IEEE utilizados para auditorías y licencias. La concesión de licencias afectaría las compras y tendría impacto, pero requiere un acuerdo o tratado internacional con todos los países que lo ratifiquen.
6. Colocar diferentes partes de la gobernanza de las IAG bajo diferentes órganos como la FIU, la OMC y la OMPI. Es probable que algo de esto suceda, pero no sería suficiente para gobernar todas las instancias de los sistemas de la IAG.
7. TransInstitución Descentralizada Semiautónoma. Ésta podría ser la más efectiva, pero la más difícil de establecer, ya que tanto las Organizaciones Descentralizadas Semiautónomas como las TransInstituciones son conceptos nuevos.

*Extraído de “Artificial General Intelligence Issues and Opportunities”, de Jerome C. Glenn contratado por la CE para realizar aportaciones a la planificación estratégica de Horizontes 2025-27.

Apéndice C

Siglas y Nombres

Siglas

AA

AAM

AFA

AGIA

AIEA

APIC

AVIIC

CAB

CAQ

CCGG

CE

CGIA

CSTE

ED

FIG

FIU

Nombres

Aprendizaje automático

Administración de Alimentos y Medicamentos

Administración Federal de Aviación

Asociación Global sobre Inteligencia Artificial

Agencia Internacional de Energía Atómica

Arquitectura predictiva de integración conjunta

Argumentos de valor de inteligencia inteligente y consciente

Convención de Armas Biológicas

Convención de Armas Químicas

Comités Coordinadores de Gobernanza Global

Comunidad Europea

Centro para la Gobernanza de la Inteligencia Artificial

Centro de Seguridad y Tecnología Emergente

Enlace de datos o enlace de descarga

Foro Intergubernamental

Fideicomiso de Inversión Unitaria

GIECC	Grupo Intergubernamental de Expertos sobre el Cambio Climático
GPU	Generador de Programas Universales
HOPIUM	Esperanza / Opio
IACG	Inteligencia Artificial para la Ciencia y el Gobierno
IAGQ	Inteligencia Artificial General Quantum
ICC	Interfaz de Cerebro Computadora
IdeC	Internet de las Cosas
IIEE	Instituto de Ingenieros Eléctricos y Electrónicos
IIIA	Instituto de Investigación de Inteligencia Artificial
IINUIDJ	Instituto Interregional de las Naciones Unidas para Investigaciones sobre la Delincuencia y la Justicia. Internet/Protocolo de control de Transmisión/Protocolo de
INTERNET/PCT/PI	Internet
IPA	Interfaz de Programación de Aplicaciones
MAA	Modelos de Aprendizaje Automático
MLG	Modelo de lenguaje de gran tamaño
OAD	Organizaciones autónomas descentralizadas
OIEA	Organismo Internacional de Energía Atómica
OIN	Organización Internacional de Normalización
OMC	Organización Mundial de Comercio
OMPI	Organización Mundial de la Propiedad Intelectual
RGDP	Reglamento General de Protección de Datos
TNP	Tratado de no proliferación de las armas nucleares
UPG	Unidades de procesamiento gráfico