Tabla de Contenidos

Lista de "problemas fuertes" de la Inteligencia Artificial

Compilación de James Manyika para la “AI2050 Initiative”:
https://ai2050.schmidtsciences.org/hard-problems/

Estos problemas se discuten en detalle en el trabajo Ten Hard Problems in Artificial Intelligence We Must Get Right

- (1) desarrollar capacidades generales de los sistemas
- (2) garantizar el rendimiento de los sistemas de IA y sus procesos de entrenamiento
- (3) alinear los objetivos del sistema con los objetivos humanos
- (4) hacer posible grandes aplicaciones de la IA para la vida real
- (5) abordar las perturbaciones económicas
- (6) asegurar la participación de todos
- (7) garantizar al mismo tiempo un despliegue socialmente responsable
- (8) abordar cualquier perturbación geopolítica que cause la IA
- (9) promover una buena gobernanza de la tecnología
- (10) gestionar las disrupciones filosóficas para los humanos que viven en la era de la IA

Acerca de la lista de problemas difíciles en IA

Basándose en trabajos previos en IA y a través de numerosas conversaciones con otros expertos, la iniciativa ha desarrollado una lista de trabajo inicial de los problemas difíciles que debe abordar AI2050. Esta lista tiene como objetivo aprovechar la oportunidad que la IA ofrece a la sociedad y abordar los riesgos y desafíos que podrían derivarse de ella.

Si bien creemos que las oportunidades y desafíos descritos en la lista de trabajo son multidisciplinarios, generalmente están dirigidos a problemas científicos y técnicos difíciles y desafíos sociales de diferentes tipos que representan tanto oportunidades como desafíos. La lista apunta a categorías relativamente distintas de desafíos y oportunidades para resolver.

Esta lista de trabajo no pretende ser exhaustiva, definitiva ni fijada en el tiempo. Esperamos que dicha lista continúe evolucionando a medida que aprendamos más y que las capacidades de la IA progresen y nuestro uso de la misma continúe evolucionando. Planeamos actualizar esta lista con el tiempo, revisando las categorías actuales, incluidas las subcategorías, y potencialmente introduciendo nuevas categorías de problemas difíciles de resolver guiados por la pregunta motivadora.

La lista de trabajo de problemas difíciles de AI2050 se compiló a partir de investigaciones y otras iniciativas en las que han participado los copresidentes Eric Schmidt y James Manyika, y aportes de numerosas conversaciones con personas a la vanguardia de la investigación y el desarrollo de la IA, y aquellos que investigan sus impactos. sobre la sociedad.

A lo largo de esta lista, “Resolver” debe entenderse como resolver o hacer avances dramáticos o lograr avances suficientes para mantenerse significativamente por delante de los desafíos o problemas emergentes a medida que la IA continúa avanzando y también evoluciona la forma en que

la sociedad y sus actores la usan o hacen mal uso de ella. .

Lista actualizada a junio de 2023

Desarrollar una IA más capaz y más general, que sea útil, segura y gane la confianza del público.

1

Resolver las limitaciones científicas y tecnológicas y los problemas difíciles de la IA actual que son fundamentales para permitir mayores avances en la IA que conduzcan a una IA más poderosa y útil capaz de aprovechar las posibilidades beneficiosas y emocionantes, incluida la inteligencia artificial general (AGI).

Los ejemplos incluyen generalizabilidad, razonamiento causal, cognición de nivel superior/meta, sistemas multiagente, cognición de agentes, la capacidad de generar nuevos conocimientos, conjeturas/teorías científicas novedosas, capacidades beneficiosas novedosas y arquitecturas informáticas novedosas, avances en el uso de recursos por parte de la IA.

2

Resolver los desafíos en constante evolución de la seguridad, la solidez, el rendimiento, la producción y otras deficiencias de la IA que pueden causar daño o erosionar la confianza pública en los sistemas de IA, especialmente en aplicaciones y usos críticos para la seguridad donde los riesgos sociales y el potencial de daño social son altos.

Los ejemplos incluyen sesgo y equidad, toxicidad de los resultados, factibilidad/exactitud, peligros de la información, incluida la desinformación, confiabilidad, seguridad, privacidad e integridad de los datos, mala aplicación, inteligibilidad y explicabilidad, daños sociales y psicológicos.

3

Resolver desafíos de seguridad y control, alineación humana y compatibilidad con una IA cada vez más poderosa y capaz y, eventualmente, AGI.

Los ejemplos incluyen riesgos asociados con el uso de herramientas/conexiones a sistemas físicos, sistemas de múltiples agentes, especificación errónea de objetivos/desviación/corrupción, riesgos de sistemas de automejora/autorreescritura, riesgos de ganancia de función y riesgos catastróficos, alineación, sistemas demostrablemente beneficiosos, cooperación hombre-máquina, desafíos de normatividad y plasticidad.

Aprovechar la IA para abordar los mayores desafíos de la

humanidad y ofrecer beneficios positivos para todos

4

Hacer contribuciones revolucionarias que permitan a la IA abordar uno o más de los mayores desafíos y oportunidades de la humanidad.

Los ejemplos incluyen los campos de las ciencias de la salud y la vida, el clima y la sostenibilidad, el bienestar humano, las ciencias fundamentales (incluidas las ciencias sociales) y las matemáticas, la exploración espacial, los descubrimientos científicos, los desafíos sociales apremiantes (por ejemplo, los Objetivos de Desarrollo Sostenible), etc.

5

Resolver los desafíos y oportunidades económicas resultantes de la IA y sus tecnologías relacionadas.

Los ejemplos incluyen nuevos modos de abundancia, escasez y uso de recursos, inclusión económica, futuro del trabajo, propiedad intelectual y creación de contenido, modelos comerciales responsables, efectos de red y competencia, y con especial atención a los países, organizaciones, comunidades y personas que no lo son. liderando el desarrollo o uso directo de la IA.

6

Resolver el acceso, la participación y la agencia en el desarrollo de la IA y el crecimiento de su ecosistema y su uso beneficioso para países, empresas, organizaciones y segmentos de la sociedad y las personas, especialmente aquellos que no están involucrados en el desarrollo de la IA.

Los ejemplos incluyen el acceso a la investigación y los recursos para el desarrollo de la IA, la diversidad de participación en el ecosistema de la IA, el acceso equitativo a capacidades y beneficios, y la diversidad disciplinaria en el desarrollo de la IA.

Desarrollar, implementar, utilizar y competir por la IA de manera responsable

7

Resolver riesgos relacionados con la IA, uso y mal uso: competición, cooperación y coordinación entre países, empresas y otros actores clave, dados los riesgos económicos, geopolíticos y de seguridad nacional.

Los ejemplos incluyen la ciberseguridad de los sistemas de IA, la gobernanza de los sistemas fronterizos/más capaces, los enfoques para controlar el uso indebido por parte de diferentes tipos de actores, la gobernanza de las armas autónomas, evitar las condiciones de carrera para el desarrollo/despliegue de la IA a expensas de la seguridad, protocolos y tratados de IA verificables y gobernar de manera estable el surgimiento de AGI.

8

Resolver los desafíos y las complejidades de la investigación responsable, el despliegue y la integración sociotécnica de la IA en diferentes dominios de uso, espacios sociales y teniendo en cuenta diferentes culturas, participantes, intereses, riesgos, externalidades y riesgos sociales, y fuerzas del mercado y de otro tipo.

Los ejemplos incluyen publicación, enfoques responsables de código abierto, distribuciones y acceso a herramientas y conjuntos de datos, enfoques de prueba/aprendizaje/iteración, enfoques relevantes para el dominio y uso y consumo de recursos responsable.

Los sistemas sociales coevolucionan y lo que significa ser humano en la era de la IA

9

Resolver la adaptación, la coevolución y la resiliencia de las instituciones de gobernanza humana y la infraestructura social y la capacidad para mantenerse al día y aprovechar el progreso de la IA en beneficio de la sociedad.

Los ejemplos incluyen la comprensión de la IA por parte de los líderes en políticas, regulación, implementación y adaptación de sistemas sociopolíticos, instituciones e infraestructuras cívicas y de gobernanza, educación y otras capacidades y sistemas humanos para permitir el florecimiento humano y social junto con una IA cada vez más capaz.

10

Resolver lo que significa ser humano en la era de la IA, o el problema señalado por John Maynard Keynes cuando dijo: “Así, por primera vez desde su creación, el hombre se enfrentará a su problema real y permanente: cómo utilizar su libertad de preocupaciones económicas apremiantes, cómo ocupar el tiempo libre que la ciencia y el interés compuesto le habrán ganado, para vivir sabiamente, agradablemente y bien”.

Los ejemplos incluyen la ética humanista junto con la poderosa IA, un mundo sin esfuerzos económicos, excepcionalismo, significado y propósito humanos.