Los 23 principios de la Conferencia Asilomar

para un uso beneficioso de la Inteligencia Artificial

La inteligencia artificial ya ha proporcionado herramientas beneficiosas que son utilizadas diariamente por personas de todo el mundo. Su continuo desarrollo, guiado por los siguientes principios, ofrecerá increíbles oportunidades para ayudar y empoderar a las personas en las décadas y siglos venideros.

Temas de investigación

1) Objetivo de la investigación: El objetivo de la investigación de la IA debe ser crear no inteligencia no dirigida, sino inteligencia beneficiosa.

2) Financiación de la investigación: Las inversiones en AI deben ir acompañadas de fondos para la investigación que asegure su uso beneficioso, incluyendo cuestiones espinosas en ciencias de la computación, economía, derecho, ética y estudios sociales, tales como:

¿Cómo podemos hacer que los futuros sistemas de IA sean altamente robustos, para que hagan lo que queramos sin que funcionen mal o sean pirateados?
¿Cómo podemos aumentar nuestra prosperidad a través de la automatización mientras mantenemos los recursos y el propósito de las personas?
¿Cómo podemos actualizar nuestros sistemas legales para ser más justos y eficientes, para mantener el ritmo de la IA y para manejar los riesgos asociados con la IA?
¿Con qué conjunto de valores debería alinearse la IA y qué estatus legal y ético debería tener?

3) Vínculo entre ciencia y política: Debe haber un intercambio constructivo y saludable entre los investigadores de la IA y los responsables de la formulación de políticas.

4) Cultura de la investigación: Se debe fomentar una cultura de cooperación, confianza y transparencia entre los investigadores y desarrolladores de la IA.

5) Evasión de razas: Los equipos que desarrollen sistemas de IA deben cooperar activamente para evitar recortes en las normas de seguridad.

Ética y Valores

6) Seguridad: Los sistemas de IA deben ser seguros durante toda su vida útil, y verificables siempre que sea posible y factible.

7) Fallo Transparencia: Si un sistema de IA causa daño, debería ser posible determinar por qué.

8) Transparencia Judicial: Toda participación de un sistema autónomo en la toma de decisiones judiciales debe proporcionar una explicación satisfactoria que pueda ser auditada por una autoridad humana competente.

9) Responsabilidad: Los diseñadores y constructores de sistemas avanzados de IA son partes interesadas en las implicaciones morales de su uso, mal uso y acciones, con la responsabilidad y oportunidad de dar forma a esas implicaciones.

10) Alineación de valores: Los sistemas de IA altamente autónomos deben ser diseñados de manera que sus objetivos y comportamientos puedan ser asegurados para alinearse con los valores humanos a lo largo de su operación.

11) Valores Humanos: Los sistemas de IA deben ser diseñados y operados de manera que sean compatibles con los ideales de dignidad humana, derechos, libertades y diversidad cultural.

12) Privacidad personal: Las personas deben tener el derecho de acceder, administrar y controlar los datos que generan, dado el poder de los sistemas de IA para analizar y utilizar esos datos.

13) Libertad y Privacidad: La aplicación de la IA a los datos personales no debe restringir irrazonablemente la libertad real o percibida de las personas.

14) Beneficio Compartido: Las tecnologías de IA deben beneficiar y empoderar a tantas personas como sea posible.

15) Prosperidad compartida: La prosperidad económica creada por AI debe compartirse ampliamente, para beneficiar a toda la humanidad.

16) Control Humano: Los humanos deben elegir cómo y si delegar decisiones a los sistemas de IA, para lograr los objetivos elegidos por el ser humano.

17) No subversión: El poder que confiere el control de sistemas de inteligencia artificial muy avanzados debe respetar y mejorar, en lugar de subvertir, los procesos sociales y cívicos de los que depende la salud de la sociedad.

18) Carrera armamentista AI: Debe evitarse una carrera armamentista con armas autónomas letales.

Temas a largo plazo

19) Capacidad Precaución: Al no haber consenso, debemos evitar las suposiciones firmes con respecto a los límites máximos de las futuras capacidades de IA.

20) Importancia: La IA avanzada podría representar un cambio profundo en la historia de la vida en la Tierra, y debería planificarse y gestionarse con el cuidado y los recursos adecuados.

21) Riesgos: Los riesgos que plantean los sistemas de IA, especialmente los riesgos catastróficos o existenciales, deben estar sujetos a esfuerzos de planificación y mitigación acordes con su impacto esperado.

22) Auto-mejora Recursiva: Los sistemas de inseminación artificial diseñados para mejorar o replicarse recursivamente de manera que puedan conducir a un rápido aumento de la calidad o la cantidad deben estar sujetos a estrictas medidas de seguridad y control.

23) Bien Común: La Superinteligencia sólo debe desarrollarse al servicio de ideales éticos ampliamente compartidos, y en beneficio de toda la humanidad y no de un solo Estado u organización.