

Primera reunión de la Red Internacional de Institutos de Seguridad de la IA

Los institutos de seguridad de la IA lanzaron la Red Internacional de Institutos de Seguridad de la IA en San Francisco. La declaración de misión refleja sus objetivos para avanzar en la seguridad, la investigación, las pruebas y la orientación de la IA.

Los días 20 y 21 de noviembre de 2024, los institutos de seguridad de la IA y las oficinas gubernamentales de Australia, el Canadá, los Estados Unidos, Francia, el Japón, Kenya, el Reino Unido, Francia, el Japón y Kenya, el Reino Unido, la República de Corea y la República de Corea se reunirán en San Francisco para celebrar la primera reunión de la Red Internacional de Institutos de Seguridad de la IA.

Sobre la base de la Declaración de Intenciones de Seúl para la Cooperación Internacional en la Ciencia de la Seguridad de la IA, publicada durante la Cumbre de la IA de Seúl el 21 de mayo de 2024, esta iniciativa marca el comienzo de una nueva fase de colaboración internacional en materia de seguridad de la IA.

La Red reúne a organizaciones técnicas dedicadas a promover la seguridad de la IA, ayudar a los gobiernos y a las sociedades a comprender mejor los riesgos que plantean los sistemas avanzados de IA y proponer soluciones para mitigar estos riesgos. Los miembros de la Red también subrayan en su declaración de misión que "la cooperación internacional para promover la seguridad, la inclusión, la inclusión y la confianza de la IA es vital para abordar estos riesgos, impulsar la innovación responsable y ampliar el acceso a los beneficios de la IA en todo el mundo".

Más allá de abordar los posibles daños, los institutos y oficinas involucrados guiarán el desarrollo y la implementación responsables de los sistemas de IA.

Objetivos y prioridades de la red

La Red Internacional de Institutos de Seguridad de la IA servirá de foro de colaboración, reuniendo conocimientos técnicos para abordar los riesgos de seguridad de la IA y las mejores prácticas. Reconociendo la importancia de la diversidad cultural y lingüística, la Red trabajará para lograr una comprensión unificada de los riesgos de seguridad de la IA y las estrategias de mitigación.

Se centrará en cuatro áreas prioritarias:

- Investigación: Colaborar con la comunidad científica para avanzar en la investigación sobre los riesgos y las capacidades de los sistemas avanzados de IA, al tiempo que se comparten los hallazgos clave para fortalecer la ciencia de la seguridad de la IA.
- Pruebas: Desarrollar y compartir las mejores prácticas para probar sistemas avanzados de IA, incluida la realización de ejercicios de prueba conjuntos y el intercambio de ideas de evaluaciones nacionales, según corresponda.
- Orientación: Facilitar enfoques compartidos para interpretar los resultados de las pruebas de los sistemas avanzados de IA a fin de garantizar respuestas coherentes y eficaces.
- Inclusión: Involucrar a los socios y partes interesadas de las regiones en todas las etapas de desarrollo, compartiendo información y herramientas técnicas de manera accesible para ampliar la participación en la ciencia de la seguridad de la IA.

Un compromiso con la cooperación global

A través de esta Red, los miembros se comprometen a promover la alineación internacional en la investigación, las pruebas y la orientación en materia de seguridad de la IA. Al fomentar la colaboración técnica y la inclusión, su objetivo es garantizar que los beneficios de una innovación en IA segura, protegida y fiable se compartan ampliamente, permitiendo a la humanidad aprovechar plenamente el potencial de la IA.

Fuente: https://digital-strategy.ec.europa.eu/

LINK DE LA NOTICIA