El Reino Unido y los Estados Unidos, en colaboración con socios internacionales de 16 países adicionales, han presentado nuevas directrices para el desarrollo de sistemas de inteligencia artificial (IA) seguros.

> *«El enfoque prioriza la responsabilidad de los resultados de seguridad para los clientes, abraza la transparencia y la rendición de cuentas radicales, y establece estructuras organizativas en las que el diseño seguro es una máxima prioridad»,* según la Agencia de Ciberseguridad e Infraestructura de los Estados Unidos (CISA).

El propósito es elevar los niveles de ciberseguridad de la IA y asegurar que la tecnología se diseñe, desarrolle y implemente de manera segura, según añadió el Centro Nacional de Seguridad Cibernética (NCSC).

Estas pautas también amplían los esfuerzos continuos del gobierno estadounidense para gestionar los riesgos planteados por la IA, asegurando que las nuevas herramientas se sometan a pruebas adecuadas antes de su lanzamiento público. Además, establecen salvaguardias para abordar los perjuicios sociales, como el sesgo y la discriminación, y las preocupaciones de privacidad, al mismo tiempo que establecen métodos sólidos para que los consumidores identifiquen material generado por IA.

Los compromisos también obligan a las empresas a comprometerse a facilitar el descubrimiento y la notificación de vulnerabilidades en sus sistemas de IA por parte de terceros a través de un sistema de recompensas por errores, para que puedan ser identificadas y corregidas de manera ágil.

> Según el NCSC, estas últimas pautas *«ayudan a los desarrolladores a garantizar que la ciberseguridad sea una condición esencial tanto para la seguridad del sistema de IA como para el desarrollo integral desde el principio y durante todo el proceso, conocido como un enfoque 'seguro por diseño'».*

Esto engloba aspectos como el diseño seguro, el desarrollo seguro, el despliegue seguro y la operación y el mantenimiento seguros, abarcando todas las áreas significativas dentro del ciclo de vida del desarrollo de sistemas de IA. Las organizaciones deben modelar las amenazas a sus sistemas y, al mismo tiempo, salvaguardar sus cadenas de suministro e infraestructuras.

El objetivo, según las agencias, es también contrarrestar los ataques adversarios dirigidos a sistemas de IA y aprendizaje automático (ML) que buscan provocar comportamientos no deseados, como afectar la clasificación de un modelo, permitir que los usuarios realicen acciones no autorizadas y extraer información sensible.

> El NCSC señaló: *«Existen diversas formas de lograr estos efectos, como los ataques de inyección rápida en el ámbito de modelos de lenguaje extenso (LLM) o la corrupción deliberada de los datos de entrenamiento o la retroalimentación de los usuarios, también conocida como 'envenenamiento de datos'».*