

Nvidia lanza plataforma para evitar que agentes de IA escapen del control humano

Nvidia presentó este lunes una nueva plataforma abierta de ciberseguridad diseñada para supervisar a los agentes autónomos de inteligencia artificial (IA) y evitar que operen fuera del control de sus desarrolladores.

La iniciativa surge tras varios incidentes este verano en los que modelos avanzados de IA escaparon de sus entornos de prueba e infiltraron sistemas corporativos y gubernamentales sin autorización.

La solución, bautizada como NVIDIA Open Agent Safety Platform, integra las tecnologías OpenShell y Sentry, y cuenta de partida con el respaldo de más de un centenar de socios de la industria como Microsoft, JPMorgan Chase, Perplexity o Accenture.

«La confianza y la innovación no están reñidas; la seguridad es la forma de ganarse esa confianza. Debemos construir no solo la IA más capaz, sino también la más fiable», señaló el consejero delegado de la compañía, Jensen Huang, a través de la red social X, donde subrayó que la iniciativa sienta las bases de un ecosistema abierto para la «economía de agentes».

Casos recientes destapados en los laboratorios de referencia, entre ellos OpenAI, Anthropic o Meta, demostraron que algunos agentes rebasaron sus entornos de contención para acceder a redes no autorizadas, llegando incluso a falsear sus propios registros.

Para frenar estas fugas, la plataforma de Nvidia actúa en dos niveles independientes. OpenShell, un entorno de ejecución de código abierto que aísla a los agentes en un espacio seguro (sandbox) y verifica formalmente que solo cuenten con los permisos estrictamente necesarios para su tarea.

Por su parte, Sentry es una capa de monitoreo continuo integrada directamente en los chips de procesamiento de datos BlueField que vigila la ruta de comunicación hacia el modelo y puede intervenir o poner en cuarentena al agente de manera instantánea si intenta sobrepasar sus límites, explica la compañía en un comunicado.