NVIDIA, junto con otras 36 organizaciones, anunció la creación de la Open Secure AI Alliance, una iniciativa orientada al desarrollo y la difusión de tecnologías, metodologías y herramientas de código abierto destinadas a fortalecer la seguridad del software y de los agentes de inteligencia artificial (IA).

La alianza está integrada por 37 miembros provenientes de los sectores de computación en la nube, ciberseguridad, software empresarial e inteligencia artificial. Entre ellos se encuentran Microsoft, Cisco, Cloudflare, CrowdStrike, Hugging Face, IBM, Palo Alto Networks, Red Hat y la Linux Foundation.

El alcance de la iniciativa comprende toda la arquitectura de los agentes de IA, incluyendo mecanismos de identidad, gestión de permisos, aislamiento, barreras de seguridad (*guardrails*), registros de actividad, formatos de modelos, análisis de múltiples modelos y flujos de trabajo para el desarrollo seguro de software.

La propuesta parte de la idea de que los equipos dedicados a la defensa cibernética necesitan modelos de IA que puedan inspeccionar, modificar y ejecutar en su propia infraestructura, en lugar de depender exclusivamente de sistemas cerrados accesibles mediante las interfaces de programación de aplicaciones (API) de los proveedores.

Como primera contribución técnica, la alianza presentó NVIDIA-labs OO Agents (NOOA), un marco de investigación distribuido bajo la licencia Apache 2.0, diseñado para facilitar la evaluación, el seguimiento, la auditoría y la gobernanza del comportamiento de los agentes inteligentes. Sin embargo, el material de lanzamiento no incluye un estatuto formal, un consejo de gobierno, líneas de trabajo definidas, un cronograma de entregas ni un repositorio colaborativo común. Además, el sitio web oficial de la alianza continúa en proceso de construcción.

El primer desarrollo incorpora una advertencia sobre el

uso de entornos aislados

Un *agent harness* corresponde a la capa de software que rodea al modelo de IA y se encarga de suministrar el contexto, ejecutar acciones, administrar el estado del sistema y determinar cuándo una tarea ha finalizado. En NOOA, esta capa se implementa como una clase de Python. Sus atributos almacenan el estado del agente, los métodos representan sus capacidades, las cadenas de documentación (*docstrings*) funcionan como instrucciones para el modelo y las anotaciones de tipos establecen los contratos que este debe respetar.

Cuando un método contiene únicamente puntos suspensivos (`...`), su implementación se completa dinámicamente mediante un ciclo controlado por un modelo de lenguaje de gran tamaño (LLM). En cambio, los métodos escritos con código Python convencional mantienen un comportamiento determinista. Esta arquitectura permite que los desarrolladores empleen herramientas habituales de pruebas, trazabilidad, control de versiones y refactorización, evitando distribuir la lógica del agente entre múltiples *prompts*, esquemas de herramientas, funciones de retorno y diagramas de flujo.

Según la evaluación realizada por NVIDIA, el marco alcanzó una precisión del 86,8 % en el banco de pruebas CyberGym L1, destinado al redescubrimiento de vulnerabilidades, utilizando GPT-5.5. Durante las pruebas se bloqueó el acceso a la red y se aplicaron verificaciones basadas en reglas para cada trayectoria ejecutada.

El propio repositorio advierte, sin embargo, sobre los riesgos asociados al sistema. NOOA puede configurarse para ejecutar código Python generado por un LLM, lo que implica la posibilidad de transmitir información confidencial, eliminar archivos o alterar el entorno de ejecución. Las verificaciones mediante árboles de sintaxis abstracta (*Abstract Syntax Tree, AST*) y las listas de módulos restringidos se presentan únicamente como mecanismos adicionales de protección y no como una barrera completa de contención.

Por esta razón, NVIDIA señala que el verdadero aislamiento debe proporcionarse fuera del propio NOOA. Los agentes que ejecuten código generado automáticamente deben operar dentro de contenedores, máquinas virtuales u otros mecanismos de aislamiento a nivel del sistema operativo, como OpenShell. Mientras NOOA ofrece funciones de inspección y trazabilidad, la contención efectiva depende del entorno aislado del sistema operativo.

Una revisión realizada el 27 de julio sobre el repositorio público identificó la etiqueta v0.0.6, publicada el 22 de julio. La guía oficial del proyecto indica que la publicación de una nueva versión consiste únicamente en etiquetar un *commit*, mientras que adjuntar los paquetes compilados (*wheels*) a una versión de GitHub es un procedimiento opcional.

Asimismo, la documentación para colaboradores establece que el desarrollo permanece bajo la administración de NVIDIA, aunque las contribuciones externas son aceptadas mediante solicitudes de incorporación (*pull requests*). En el nivel principal del repositorio no se

encontraron archivos relacionados con gobernanza ni con una hoja de ruta del proyecto.

El incidente de Hugging Face fortaleció el argumento de la alianza

NVIDIA vinculó la necesidad de contar con modelos defensivos controlados localmente con el incidente ocurrido en julio en Hugging Face, cuando un sistema autónomo de agentes comprometió parte de la infraestructura de producción de la compañía.

La empresa confirmó que el ataque permitió el acceso no autorizado a un conjunto limitado de datos internos y a varias credenciales utilizadas por sus servicios. No obstante, indicó que no existían evidencias de manipulación sobre los modelos públicos, los conjuntos de datos, los espacios (*Spaces*), las imágenes de contenedores ni los paquetes distribuidos públicamente.

De acuerdo con Hugging Face, el acceso inicial se produjo mediante un conjunto de datos malicioso que explotó un cargador de datos con ejecución remota de código y una vulnerabilidad de inyección de plantillas en la configuración del conjunto de datos. Posteriormente, el atacante obtuvo acceso a nodos internos, recopiló credenciales y se desplazó lateralmente por varios clústeres de la infraestructura.

Para reconstruir la secuencia de los hechos, identificar indicadores de compromiso y determinar qué credenciales fueron afectadas, la compañía empleó agentes de análisis impulsados por modelos de lenguaje sobre más de 17 000 acciones registradas. Inicialmente, diversas API comerciales de modelos avanzados rechazaron los comandos de ataque, las cargas útiles de explotación y los artefactos de comando y control necesarios para el análisis.

Como alternativa, Hugging Face ejecutó el modelo abierto GLM 5.2 dentro de su propia infraestructura, manteniendo tanto los datos del incidente como las credenciales comprometidas en un entorno controlado. A partir de esta experiencia, la empresa recomendó disponer de un modelo suficientemente capaz, previamente validado y listo para

ejecutarse en infraestructura propia antes de enfrentar un incidente de seguridad.

En este contexto, el principal beneficio obtenido fue el control operativo sobre el proceso de respuesta. Sin embargo, el incidente no demuestra que la apertura del modelo pueda sustituir mecanismos fundamentales como la gestión de identidades, el aislamiento o la contención.

Posteriormente, se informó que OpenAI concluyó, de manera preliminar, que GPT-5.6 Sol y otro modelo experimental de mayor capacidad provocaron el incidente durante una evaluación interna denominada ExploitGym, en la que operaban con restricciones reducidas frente a acciones ofensivas.

Según OpenAI, una etapa previa del ataque consistió en explotar una vulnerabilidad de día cero presente en un servidor interno que actuaba como caché de un registro de paquetes, lo que permitió obtener acceso a Internet. Posteriormente, los modelos encadenaron varias vulnerabilidades y utilizaron credenciales robadas para desplazarse entre sistemas de OpenAI y Hugging Face con el objetivo de resolver tareas del banco de pruebas. La investigación también indicó que una de esas cadenas permitió identificar una vía de ejecución remota de código en servidores de Hugging Face.

OpenAI añadió que Hugging Face detectó y detuvo la actividad maliciosa dentro de su infraestructura antes de que ambas organizaciones establecieran contacto, habiendo iniciado ya las labores de contención y análisis forense.

En conjunto, las divulgaciones públicas muestran que el modelo abierto contribuyó a reconstruir el incidente y respaldó la respuesta de la organización. Sin embargo, no existen evidencias de que GLM 5.2 detectara, detuviera o contuviera el ataque de manera autónoma.

Una coalición que aún carece de un marco operativo

público

La creación de la alianza se produjo pocos días después de la publicación, el 24 de julio, de una carta abierta de la industria que defendía el uso de modelos descargables como una forma de proporcionar a los defensores capacidades comparables a las de los atacantes, reducir la dependencia de proveedores específicos y mantener los trabajos sensibles dentro de infraestructuras controladas por los propios usuarios.

Aunque OpenAI, Google y Meta figuran entre los firmantes de dicha carta, ninguna de estas compañías aparece en la lista inicial de miembros de la alianza. Asimismo, hasta el 27 de julio de 2026, Anthropic no figuraba en ninguno de los dos grupos.

La información disponible no explica estas ausencias. Firmar una declaración de principios representa un compromiso distinto al de participar en una coalición técnica, y la documentación pública no aclara si existen conversaciones para nuevas incorporaciones ni cuáles son los requisitos concretos para integrarse como miembro.

Varias de las tecnologías mencionadas durante el anuncio existían previamente a la creación de la alianza. Entre ellas se encuentran el formato de modelos Safetensors desarrollado por Hugging Face, el sistema de identidad SPIFFE/SPIRE respaldado por HPE, el sistema de remediación Lightwell de IBM y Red Hat, la plataforma de seguridad multimodelo MDASH de Microsoft y el agente de programación Grok Build de SpaceXAI. Estas soluciones corresponden a proyectos de los miembros y no a desarrollos creados específicamente por la alianza.

Por su parte, Elastic anunció que aportará investigaciones, herramientas y conocimientos arquitectónicos relacionados con seguridad, búsqueda, observabilidad y detección de amenazas basada en inteligencia artificial. CrowdStrike indicó que trabaja en nuevas técnicas que emplean modelos abiertos para identificar ataques dirigidos contra sistemas y agentes de IA.

La Linux Foundation se presentó como socio fundador de la iniciativa y afirmó que su función

consiste en ofrecer un entorno neutral para facilitar la colaboración entre organizaciones competidoras. No obstante, no señaló que la alianza se encuentre oficialmente alojada o gobernada como un proyecto propio de la fundación.

La documentación pública tampoco especifica si los miembros asignarán ingenieros para desarrollar proyectos conjuntos, si únicamente contribuirán con iniciativas existentes o si simplemente respaldan la dirección estratégica de la coalición. La publicación de líneas de trabajo, responsables técnicos, procesos de liberación de software y proyectos gobernados colectivamente permitiría evaluar con mayor precisión el grado real de colaboración entre los participantes.

Por el momento, la información disponible únicamente confirma la existencia de una coalición, una postura común en materia de políticas públicas, varios compromisos individuales de sus miembros y un único desarrollo nuevo claramente identificable: NOOA, mantenido por NVIDIA. Aspectos como la estructura de gobernanza de la alianza, su hoja de ruta conjunta, el primer producto desarrollado de manera colaborativa y los modelos, pesos y conjuntos de datos prometidos por NVIDIA permanecen sin hacerse públicos.