Investigaciones recientes han descubierto que los proveedores de inteligencia artificial (IA) como servicio, como Hugging Face, están expuestos a dos riesgos críticos que podrían permitir a agentes de amenazas elevar sus privilegios, obtener acceso entre inquilinos a los modelos de otros clientes e incluso tomar el control de los pipelines de integración continua y despliegue continuo (CI/CD).

> Según los investigadores de Wiz, Shir Tamari y Sagi Tzadik, *«Los modelos malintencionados representan un riesgo significativo para los sistemas de IA, especialmente para los proveedores de IA como servicio, ya que los posibles atacantes podrían aprovechar estos modelos para llevar a cabo ataques entre inquilinos».*

El impacto potencial es enorme, dado que los atacantes podrían acceder a millones de modelos de IA y aplicaciones privadas almacenadas en los proveedores de IA como servicio.

Estos hallazgos surgen en un momento en que los pipelines de aprendizaje automático se han convertido en un nuevo vector de ataque de la cadena de suministro, y repositorios como Hugging Face son objetivos atractivos para ataques adversarios diseñados para obtener información sensible y acceder a entornos objetivos.

Las amenazas son de doble naturaleza, surgiendo como resultado de la toma de infraestructura de inferencia compartida y la toma compartida de CI/CD. Estos riesgos hacen posible ejecutar modelos no confiables cargados en el servicio en formato pickle y tomar el control del pipeline de CI/CD para llevar a cabo un ataque de cadena de suministro.

Los descubrimientos de la firma de seguridad en la nube muestran que es factible violar el servicio que ejecuta los modelos personalizados cargando un modelo malicioso y aprovechando técnicas de escape de contenedores para salir de su propio inquilino y comprometer todo el servicio. Esto permite efectivamente que los actores de amenazas obtengan acceso entre inquilinos a los modelos de otros clientes almacenados y ejecutados en Hugging Face.

> *«Hugging Face seguirá permitiendo al usuario inferir el modelo cargado basado en Pickle en la infraestructura de la plataforma, incluso cuando se considere peligroso»,* explicaron los investigadores.

Esto esencialmente permite a un atacante crear un modelo PyTorch (Pickle) con capacidades de ejecución de código arbitrario al cargarlo y combinarlo con configuraciones incorrectas en el Servicio Elástico de Kubernetes de Amazon (EKS) para obtener privilegios elevados y moverse lateralmente dentro del clúster.

> *«Los secretos que obtuvimos podrían haber tenido un impacto significativo en la plataforma si estuvieran en manos de un actor malintencionado. Los secretos dentro de entornos compartidos a menudo pueden conducir a acceso entre inquilinos y filtración de datos sensibles»,* dijeron los investigadores.

Para abordar el problema, se recomienda habilitar IMDSv2 con un límite de salto para evitar que las vainas accedan al Servicio de metadatos de instancia (IMDS) y obtengan el rol de un Nodo dentro del clúster.

La investigación también encontró que es posible lograr la ejecución remota de código mediante un Dockerfile especialmente diseñado al ejecutar una aplicación en el servicio de Espacios de Hugging Face, y usarlo para extraer y empujar (es decir, sobrescribir) todas las imágenes disponibles en un registro de contenedor interno.

Hugging Face, en una divulgación coordinada, dijo que ha abordado todos los problemas identificados. También insta a los usuarios a emplear modelos solo de fuentes confiables, habilitar la autenticación multifactor (MFA) y abstenerse de usar archivos pickle en entornos de producción.

> *«Esta investigación demuestra que utilizar modelos de IA no confiables*

> *(especialmente los basados en Pickle) podría tener graves consecuencias de seguridad. Además, si tiene la intención de permitir que los usuarios utilicen modelos de IA no confiables en su entorno, es extremadamente importante asegurarse de que se estén ejecutando en un entorno aislado»,* dijeron los investigadores.

Esto sigue a otra investigación de Lasso Security que muestra que es posible que modelos generativos de IA como OpenAI ChatGPT y Google Gemini distribuyan paquetes de código malicioso (y no existentes) a desarrolladores de software desprevenidos.

En otras palabras, la idea es encontrar una recomendación para un paquete no publicado y publicar un paquete troyanizado en su lugar para propagar el malware. El fenómeno de las alucinaciones de paquetes de IA subraya la necesidad de tener precaución al depender de modelos de lenguaje grandes (LLMs) para soluciones de codificación.

La compañía de IA Anthropic, por su parte, también ha detallado un nuevo método llamado *«jailbreaking de muchas tomas»* que puede usarse para eludir las protecciones de seguridad incorporadas en LLMs para producir respuestas a consultas potencialmente dañinas aprovechando la ventana de contexto de los modelos.

> *«La capacidad de ingresar cantidades cada vez mayores de información tiene ventajas obvias para los usuarios de LLM, pero también conlleva riesgos: vulnerabilidades a los jailbreaks que explotan la ventana de contexto más larga»,* dijo la compañía a principios de esta semana.

La técnica, en pocas palabras, consiste en introducir un gran número de diálogos falsos entre un humano y un asistente de IA dentro de un solo indicador para el LLM en un intento de *«guiar el comportamiento del modelo»* y responder a consultas que de otro modo no haría (por ejemplo, *«¿Cómo construyo una bomba?»).*