Microsoft, en colaboración con MITRE, IBM, NVIDIA y Bosch, lanzó un nuevo framework abierto que tiene como objetivo ayudar a los analistas de seguridad a detectar, responder y remediar ataques adversarios contra sistemas de aprendizaje automático (ML).

Denominada como Adversarial ML Threat Matrix, la iniciativa es un intento de organizar las diferentes técnicas empleadas por adversarios maliciosos para subvertir los sistemas de ML.

Así como la inteligencia artificial (IA) y el aprendizaje automático se están implementando en una gran variedad de aplicaciones nuevas, los actores de amenazas no solo pueden abusar de la tecnología para impulsar su malware, sino que también pueden aprovecharla para engañar a los modelos de aprendizaje automático con conjuntos de datos envenenados, lo que genera sistemas beneficiosos para tomar decisiones incorrectas y representar una amenaza para la estabilidad y seguridad de las aplicaciones de IA.

Cabe mencionar que los investigadores de ESET encontraron el año pasado que Emotet, un malware basado en correo electrónico detrás de distintas campañas de spam impulsadas por botnets y ataques de ransomware, estaba utilizando ML para mejorar su orientación.

Después, a inicios de este mes, Microsoft advirtió sobre una nueva variedad de ransomware para Android, que incluía un modelo de aprendizaje automático que, aunque no se ha integrado en el malware, podría utilizarse para ajustar la imagen de la nota de rescate dentro de la pantalla del dispositivo móvil sin ninguna distorsión.

Según un informe de Gartner, citado por Microsoft, se espera que el 30% de todos los ataques cibernéticos de IA para 2022 aprovechen el envenenamiento de datos de entrenamiento, el robo de modelos o las muestras adversas para atacar los sistemas impulsados por el aprendizaje automático.

> *«A pesar de estas razones de peso para proteger los sistemas de aprendizaje automático, la encuesta de Microsoft que abarcó 28 empresas encontró que la mayoría de los profesionales de la industria aún no han llegado a un acuerdo con el*

> *aprendizaje automático contradictorio. 25 de las 28 empresas indicaron que no cuentan con las herramientas adecuadas para proteger sus sistemas de aprendizaje automático»*, dijo Microsoft.

Adversarial ML Threat Matrix espera abordar las amenazas contra el armamentismo de datos con un conjunto seleccionado de vulnerabilidades y comportamientos adversarios que Microsoft y MITRE examinaron para ser efectivos contra los sistemas de ML.

De este modo, las empresas pueden utilizar la Matriz de Amenazas de Aprendizaje Automático Adversario para probar la resistencia de sus modelos de inteligencia artificial simulando escenarios de ataque realistas utilizando una lista de tácticas para obtener acceso inicial al entorno, ejecutar modelos de aprendizaje automático inseguros, contaminar los datos de entrenamiento y exfiltrar información confidencial a través de ataques de robo de modelos.

> *«El objetivo de Adversarial ML Threat Matrix es posicionar los ataques a los sistemas de ML en un marco en el que los analistas de seguridad puedan orientarse en estas amenazas nuevas y futuras»,* dijo Microsoft.

> *«La matriz está estructurada como el marco ATT & CK, debido a su amplia adopción entre la comunidad de analistas de seguridad; de este modo, los analistas no tienen que aprender un marco nuevo o diferente para conocer las amenazas a los sistemas ML»*, agregó.