Identifizierung versteckter Risiken in KI-Systemen

Manuel Tobies
IT-Security Analyst

Die NSide Attack Logic GmbH gehört zu den wenigen Anbietern, die auf offensive KI-Sicherheit setzen und Modelle wie Infrastruktur unter realen Angriffsbedingungen prüfen. Manuel Tobies erklärt, wie NSide damit Unternehmen zu mehr Resilienz und Compliance verhilft.

Herr Tobies, was genau kann man sich unter einem KI-Penetrationstest vorstellen?
Bei einem KI-Penetrationstest wird die KI-Infrastruktur des Kunden vollumfänglich auf Schwachstellen analysiert. Dies schließt nicht nur das eigentliche Machine-Learning-Modell, sondern auch dessen Einbettung in eine Anwendungsumgebung, angebundene Services und Applikationen wie auch eingesetzte Schutzmaßnahmen ein. Über diesen Ansatz können Schwachstellenketten identifiziert und die tatsächliche Ausnutzbarkeit und damit Kritikalität dieser besser eingeschätzt werden.

Ihre eigene Analyseinfrastruktur gilt als Herzstück der offensiven KI-Sicherheit bei NSide. Welche Vorteile bietet sie gegenüber Standardlösungen?
Das eigene Hosting der Analyseinfrastruktur bietet uns die Möglichkeit, den Kunden größtmögliche Datensicherheit durch vollständige Separierung der verwendeten Infrastruktur und lokale Verarbeitung ihrer Daten während des Tests garantieren zu können.

Sie trainieren eigene Modelle für Angriffsanalysen und OSINT. Welche Erkenntnisse gewinnen Sie daraus für reale Testszenarien?
Einerseits garantieren wir durch eigenes Training, Fine-Tuning, Hosting etc. ein vollständiges Verständnis der benötigten Infrastruktur, Technologien und Vorgehensweisen, andererseits bringen uns die eigenen Modelle zweckgebunden deutlich verbesserte Ergebnisse bei beispielsweise der Analyse und Detektion von Daten, die für physische Penetrationstests benötigt werden. Ein Beispiel ist hier die Detektion und Extraktion von Mitarbeiterausweisen aus Firmenvideos.

Viele Unternehmen unterschätzen KI-Risiken in bestehenden Anwendungen. Wie helfen Sie, diese Systeme sicher und compliant zu machen?
Durch Tests nach anerkannten Standards wie unter anderem der OWASP Top 10 Standard für LLM-Anwendungen werden die zu testenden Systeme auf Schwachstellen analysiert – anschließend wird ein professioneller Bericht über den Test mit allen notwendigen Details zur Schwachstellenbehebung ausgeliefert. Unterstützt durch den Bericht kann beispielsweise eine Compliance mit dem durch Google veröffentlichten Standard SAIF nachgewiesen werden oder bei Detektion schwerwiegender Schwachstellen durch Behebung dieser die Compliance hergestellt werden.

Weitere Informationen unter:
nsideattacklogic.de

NSIDE ATTACK LOGIC