VRG23-011 - Building Robust and Explainable AI-based Defenses for Computer Security (BREADS)

Zusammenfassung

Die digitale Welt sicherer machen: Künstliche Intelligenz (KI) ist in unserem Alltag allgegenwärtig und ihre Bedeutung wird durch neuartige KI-Modelle, wie OpenAI's ChatGPT oder Google's Gemini, weiter zunehmen. Diese Technologien revolutionieren viele Bereiche unseres Lebens, etwa in der Bildung und der Medizin, indem sie maßgeschneiderte Lösungen und Unterstützung bieten. Leider jedoch birgt der Fortschritt in der KI auch erhebliche Risiken: Die Fähigkeiten dieser Modelle können für schädliche Zwecke missbraucht werden, etwa zur Verbreitung von Desinformation oder zur automatischen Generierung von Schadsoftware (Malware).

In dem Projekt BREADS forscht Daniel Arp mit seinem Team an neuen KI-basierten Erkennungssystemen, um mit der sich schnell ändernden Bedrohungslage Schritt halten zu können. Hierbei werden gezielt existierende Schwächen aktueller Erkennungssysteme adressiert. Zum einen sollen durch den Einsatz erklärbarer Lernmethoden Systeme entstehen, die für deren Nutzer:innen nachvollziehbare Entscheidungen liefern, anstatt als "Black-Box" zu agieren. Zum anderen sollen sich die Erkennungssysteme automatisch an die sich kontinuierlich ändernde Bedrohungslage anpassen können. Durch die Kombination aus Erklärungen und robuster Erkennung soll es möglich werden, auch neu aufkeimenden KI-generierten Cyberbetrug, wie Deep Fakes oder automatisch generierter Malware, effektiv zu begegnen.

Wissenschaftliche Disziplinen:
IT security (60%) | Machine learning (40%)

Keywords:
computer security, machine learning, fraud detection, malware detection, dataset shift, explainable AI

VRG leader:	Daniel Arp
Institution:	TU Berlin
Proponent:	Matteo Maffei
Institution:	TU Wien

Status: Laufend (02.09.2024 - 01.09.2030)
GrantID: 10.47379/VRG23011

Weiterführende Links zu den beteiligten Personen und zum Projekt finden Sie unter
https://wwtf.at/funding/programmes/ict/VRG23-011/