Künstliche Intelligenz

Cybersecurity – Herausforderungen für die politische Schweiz

Stand der Dinge

Die Begriffe maschinelles Lernen und künstliche Intelligenz (KI) decken ein weites Feld ab, das aus der Statistik und der Operationsforschung stammt. Im Kern handelt es sich bei KI um eine Methode zur Vorhersage von Ergebnissen aus Datensätzen. Dies geschieht durch die Erstellung von Modellen, die automatisch Muster in den Daten ableiten und diese Muster zur Entscheidungsfindung nutzen. KI entwickelt sich rasch zu einer kritischen Technologie für die digitale Gesellschaft und Industrie. Wir sind zunehmend von der Fähigkeit der KI abhängig, aus früheren Erfahrungen zu lernen, Schlussfolgerungen zu ziehen, Zusammenhänge zu entdecken oder komplexe Daten zu klassifizieren, um kritische Entscheidungen zu treffen und Prozesse und Entscheidungsabläufe zu automatisieren.

Die Durchdringung der KI führt zur «Adversarial Artificial Intelligence (AAI)», bei der Angreifer (A) die KI ausnutzen, um in Gebrauch befindliche KI-Modelle zu kompromittieren, und die (B) KI nutzen, um Elemente von Angriffen zu skalieren und zu automatisieren, die vorher unmöglich waren (DeepFakes) oder stark auf manuellen Prozessen beruhten.

AAI führt dazu, dass Modelle des maschinellen Lernens Eingaben falsch interpretieren und sich in einer für den Angreifer günstigen Weise verhalten.

Empfehlungen

1. Organisationen müssen ihre KI-Modelle sowie ihre KI-gesteuerte Automatisierung und Entscheidungsfindung in die Risikobewertung einbeziehen.
2. Es muss ein Verständnis dafür entwickelt werden, wo die sich verändernden AAI-gesteuerten Bedrohungen die derzeitige Lage in Frage stellen und wo begrenzte Ressourcen konzentriert eingesetzt werden sollten.
3. Die Prüfung und Bewertung der Robustheit von KI-Modellen muss in die Entscheidungsfindung mit einbezogen werden.
4. Regierung, Wirtschaft und Bildungssysteme müssen sicherstellen, dass sie über die erforderliche Qualifikationsbasis und Talentpipeline verfügen und diese weiterentwickeln.

Um das Verhalten eines Modells zu kompromittieren, erstellen Angreifer «gegnerische Daten», die oft normalen Eingaben ähneln, aber stattdessen die Leistung des Modells beeinträchtigen. KI-Modelle klassifizieren dann diese gegnerischen Daten falsch, um mit hoher Genauigkeit inkorrekte Antworten zu liefern.

Herausforderungen

Günstige Rechenleistung und die Fülle der gesammelten Daten haben es den Modellierern und Angreifern ermöglicht, zu geringen Kosten immer komplexere KI-Modelle zu entwickeln. Diese steigende Genauigkeit und Komplexität von KI-Modellen hat dazu geführt, dass sich viele Verhaltensweisen der Modelle jedem umfassenden menschlichen Verständnis entziehen. Die meisten KI-Modelle sind so zu Black Boxes geworden. Wenn ein Angreifer ein bestimmtes Verhalten in einem KI-Modell entdeckt, das seinen Entwicklern unbekannt ist, kann er dieses Verhalten für seinen potenziellen Vorteil ausnutzen.

Verschiedene KI-Modelle, einschliesslich hochmoderner neuronaler Netze, sind anfällig für gegnerische Daten. Diese Modelle klassifizieren die Daten, die sich nur geringfügig (für Menschen unmerklich) von korrekt klassifizierten Daten unterscheiden, falsch.

Die Anfälligkeit für AAI wird zu einem der Hauptrisiken bei der Anwendung von KI in sicherheitskritischen Umgebungen. Angriffe auf Kerntechnologien wie Bildverarbeitung, optische Zeichenerkennung (OCS), Verarbeitung natürlicher Sprache (NLP), Sprache und Video (DeepFakes) und Erkennung von Malware wurden bereits nachgewiesen.

Beispiele für AAI-Bedrohungen sind:

1. Face Swapping (Gesichtertausch) in Videos unter Verwendung von maschinellen Lernalgorithmen. DeepFake-Dienste werden online bereits für wenige Dollar angeboten.
2. Bilderkennung/Klassifizierung im Bereich des autonomen Fahrens, z.B. Fehlinterpretation von Strassenschildern oder Hindernissen.
3. Manipulation der Texterkennung in automatisierten Dienstleistungen der Dokumenten- oder Zahlungsverarbeitung.
4. Fähigkeit, KI-gesteuerte Betrugserkennungs- und Kontrollmechanismen zu umgehen und zu industrialisieren.
5. Böswillige Beeinflussung von KI-Modellen zur Begünstigung oder Diskreditierung bestimmter Gruppen.

Handlungsbedarf

AAI zielt auf Bereiche der Angriffsfläche, die wir noch nie zuvor gesichert haben, die KI-Modelle selbst. Organisationen müssen ihre KI-Modelle und die KI-gesteuerte Automatisierung und Entscheidungsfindung in ihre Risikobewertung mit einbeziehen. Die Verteidigung gegen AAI umfasst proaktive und reaktive Strategien. Proaktive Strategien machen KI-Modelle robuster gegenüber gegnerischen Angriffsmustern, während reaktive Strategien darauf abzielen, gegnerische Verhaltensweisen zu erkennen, wenn das KI-Modell im Einsatz ist.

Wir müssen dieses neue und sich entwickelnde Bedrohungsumfeld erkennen und ein Verständnis dafür entwickeln. Die Prozesse, die durch automatisierte KI-Entscheidungsfindung angetrieben werden, müssen zunehmend in Frage gestellt werden.

Referenzen

Explaining and Harnessing Adversarial Examples: https://arxiv.org/pdf/1412.6572.pdf

AI Is The New Attack Surface*:* https://www.accenture.com/_acnmedia/Accenture/Redesign-Assets/DotCom/Documents/Global/1/Accenture-Trustworthy-AI-POV-Updated.pdf

What is adversarial artificial intelligence and why does it matter?: https://www.weforum.org/agenda/2018/11/what-is-adversarial-artificial-intelligence-is-and-why-does-it-matter/

Deepfakes web β: https://deepfakesweb.com

Kontakt

Nicole Wettstein
Leiterin Schwerpunktprogramm Cybersecurity
+41 44 226 50 13

https://www.satw.ch/cybersecurity-herausforderungen

Impressum
Schweizerische Akademie der Technischen Wissenschaften SATW
Expertenbeiträge
Karl Aberer, EPFL | Umberto Annino, InfoGuard | Alain Beuchat, Banque Lombard Odier & Cie SA | Matthias Bossardt, KPMG | Adolf Doerig, Doerig & Partner | Stefan Frei, ETH Zürich | Roger Halbheer, Microsoft | Pascal Lamia, MELANI | Martin Leuthold, Switch | Hannes Lubich, Verwaltungsrat und Berater | Adrian Perrig, ETH Zürich | Raphael Reischuk, Zühlke Engineering AG | Riccardo Sibilia, VBS | Bernhard Tellenbach, ZHAW | Daniel Walther, Swatch Group Services | Andreas Wespi, IBM Research Lab
Redaktion und Grafik
Beatrice Huber; Claude Naville, Adrian Sulzer, Nicole Wettstein

www.satw.ch
September 2020