Mittelstand-Digital Zentrum Chemnitz

Künstliche Intelligenz und IT-Sicherheit - Chancen

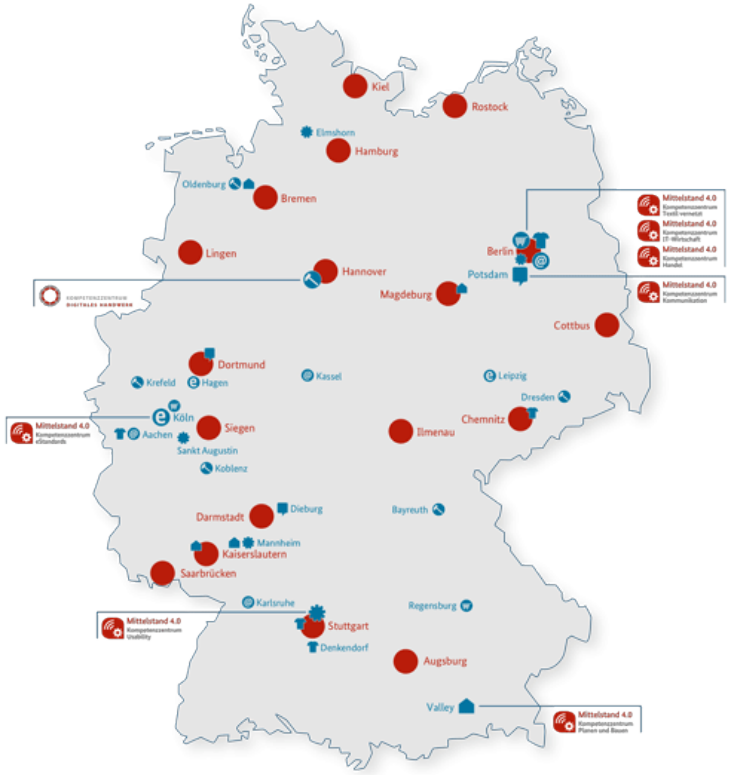
Mike Wäsche
Mittelstand-Digital Zentrum Chemnitz
c/o tti Technologietransfer und Innovationsförderung Magdeburg GmbH

Mittelstand-Digital

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Mittelstand-Digital Zentrum Chemnitz

- als Teil von Mittelstand-Digital unterstützen wir KMU und Handwerk bei der Digitalisierung
- der Mensch im Mittelpunkt -
  als Befähiger digitaler Produktions- und Arbeitswelten
- Expertise u.a. in den Bereichen Geschäftsmodell-Entwicklung, Produktion und Logistik, IT-Sicherheit und Datenschutz, Recht und Produktgestaltung



Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Was kann eine Künstliche Intelligenz bei IT-Sicherheit?

Chancen und Risiken von KI allgemein

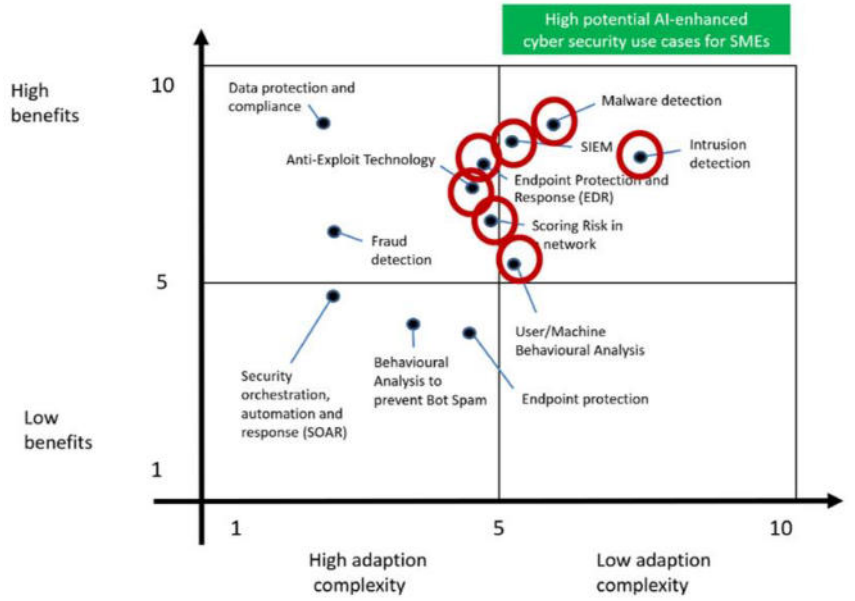
Chancen	Risiken
Effiziente Auswertung großer Datenmengen	Ausnutzen von Schwachstellen
Musteranalyse und -erkennung	Optimierung von Schadsoftware und Angriffen
Erkennung von Anomalien	Stimmen und Gesichtserkennung
Clusterung und Einordnung von SPAM-E-Mails und Malware	Beschleunigung der Kryptoanalyse
Beheben von Bugs, Automatisches Einspielen von Patches	Verschlüsselungsverfahren

Gefördert durch: Bundesministerium für Wirtschaft und Klimaschutz aufgrund eines Beschlusses des Deutschen Bundestages

Künstliche Intelligenz und IT-Sicherheit – Chancen

Geeignete Anwendungsfälle für KMU

- Malwareerkennung
- Anti Exploit Technology
- Intrusion Detection
- Endpoint Protection and Response (EDR)
- User and Entity Behavior Analytics
- Scoring Risk in a network
- Security Information and Event Management (SIEM)



Quelle: Kant, D.; Johannsen, A. (2022) Evaluation of AI-based use cases for enhancing the cybersecurity defense of smal land medium-sized companies(SMEs), in: IS&T International Symposium on Electronic Imaging 2022, Mobile Devices and Multimedia: EnablingTechnologies, Algorithms, and Applications, USA.pp 387-1 -387-8. https://doi.org/10.2352/EI.2022.34.3.MOBMU-387

Gefördert durch: Bundesministerium für Wirtschaft und Klimaschutz aufgrund eines Beschlusses des Deutschen Bundestages

Einsatzszenario

Malwareerkennung – Hintergrund

- Schadsoftware
  - Diebstahl von Daten
  - Verschlüsselung vertraulicher Daten
  - Übernahme von Systemfunktionen
  - Infektion von Geräten im Netzwerk
- unterschiedliche Formen
  - z.B. Ransomware, Spyware, Adware, Botnets oder Trojaner
- sehr große Verbreitung mit vielen Varianten

Quelle: https://b2b-cyber-security.de/wp-content/uploads/2022/12/Kaspersky-Number-of-the-Year-Infografik-DE-1024x853.webp

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Einsatzszenario

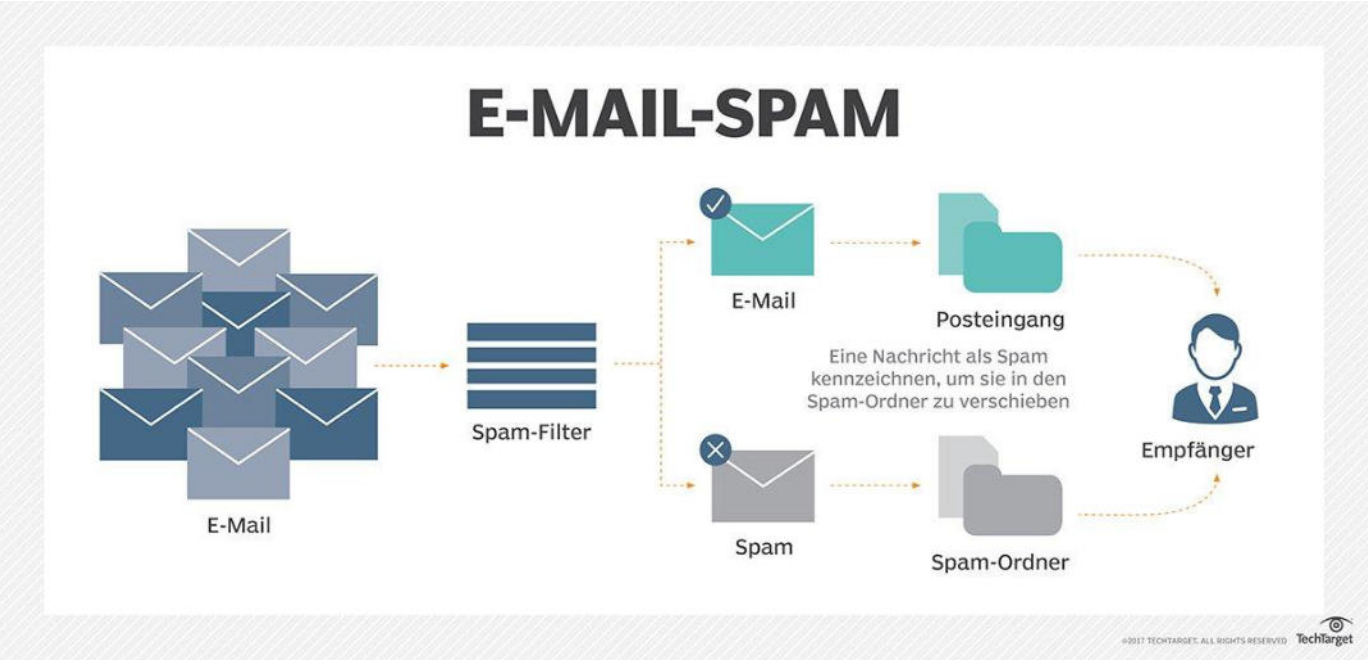
Malwareerkennung – Anwendung von KI durch Schutzsoftware

- zahlreiche verschiedene Lösungen am Markt
  - freie und kommerzieller Systeme
  - KI ist meist integriert
  - Achtung: private vs. berufliche Nutzung
- Nutzung von KI-Algorithmen
  - Vorteile gegenüber etablierten signaturbasierten Verfahren
  - Einsatz zur Mustererkennung sowie für Quelltextanalysen
- diverse Bestenlisten im Internet
  - z.B. https://www.av-test.org oder Fachzeitschriften

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Einsatzszenario

Malwareerkennung – Spamfilter



Quelle: https://www.computerweekly.com/de/definition/Spam-Filter

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Einsatzszenario

Anti Exploit Technology – Hintergrund

Quelle: https://de.fotolia.com/p/202289213

- Was ist ein Exploit?
  - systematische Methoden, eine Schwachstelle auf einem System auszunutzen
- Arten von Exploits
  - DOS-, remote-, lokale-, Zero-Day- oder Kommandozeilen-Exploits
- betroffene Anwendungen
  - Unterscheidung von anwendungsbasierten und clientspezifischen Lösungen
  - besonders gefährlich für KMU: Office-Anwendungen, Browser, PDF-Reader

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Einsatzszenario

Anti Exploit Technology – Anwendungsmöglichkeit der KI

- integrierter Schutz
  - Updates aller Anwendungen auf dem Computer aktuell halten
- Schutz durch installierte Anwendung
  - zusätzlich zum Virenscanner
  - ständige Überprüfung im Hintergrund und Möglichkeit der Behebung

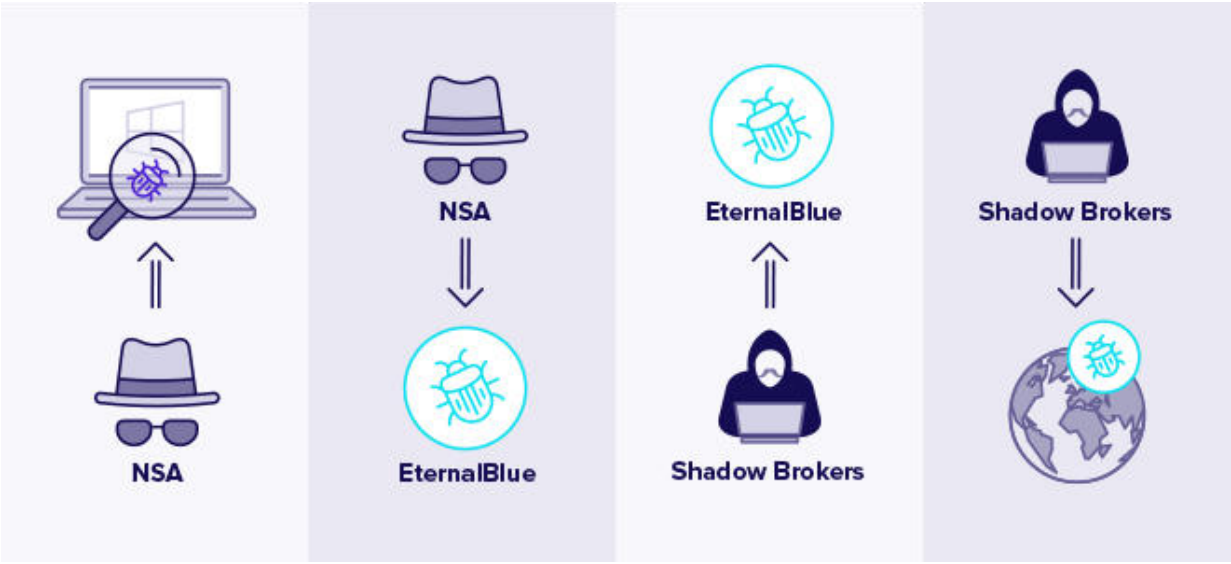
Quelle: https://www.avast.com/de-de/c-exploits

Gefördert durch: Bundesministerium für Wirtschaft und Klimaschutz aufgrund eines Beschlusses des Deutschen Bundestages

Einsatzszenario

Anti Exploit Technology – Beispiel: EternalBlue

- Entwicklung durch die NSA
- Diebstahl durch die Gruppierung Shadow Brokers
- Ausnutzung einer Schwachstelle im SMB (Server Message Block)
- Grundlage für weitere bekannte Cyberangriffe
  - WannaCry oder auch Petya



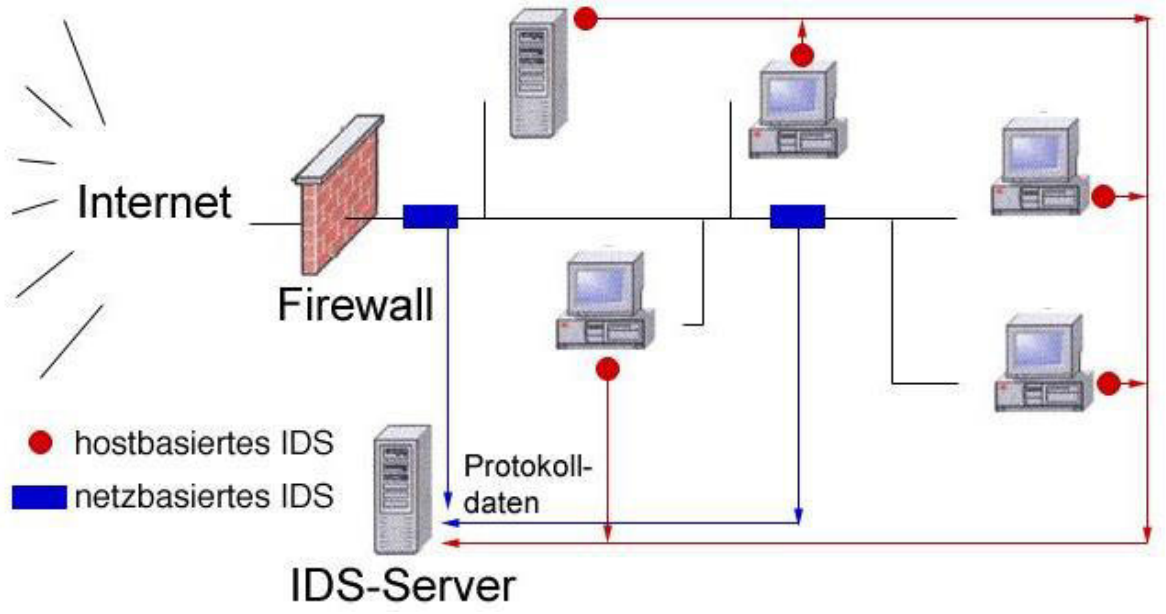
Quelle: https://www.avast.com/de-de/c-eternalblue

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Einsatzszenario

Intrusion Detection – Anwendung von KI

- aktive Überwachung der Computer-systeme
- Erkennung von Angriffen und Missbrauch
- Zusammensetzung der Werkzeuge individuell möglich
- Unterscheidung diverser Systeme möglich
  - hostbasierte, netzwerkbasierte oder hybride Systeme



Quelle: http://www.lse.de/papiere/internet%20und%20netze/netzwerk/ids.php

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Einsatzszenario

Endpoint Protection and Response (EDR) / Extendend Protection and Response (XDR)

EDR	XDR
Schutz und Abwehr von Bedrohungen gegenüber den Endgeräten mit Netzwerk	Erweiterung des Schutzes der Endgeräte um das Unternehmensnetzwerks an sich
Aufzeichnung des Verhaltens + Analyse der Daten	Aufzeichnung + Zusammenführung des Datenverkehrs im Netzwerk
verdächtiges Verhalten erkannt? – Automatisierte Reaktion zur Abwehr – Isolierung des Endgeräts	schnellere Erkennung von Unstimmigkeiten im Netzwerk
	schnellere Reaktion und Isolierung bei Angriffen

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Einsatzszenario

Scoring Risk in a Network – Hintergrund

- Bewertung der Risiken in einem Netzwerk anhand eines geeigneten Bewertungssystems (z.B. Matrix)
- Unterscheidung in externe und interne Einflussfaktoren
  - externe: Unternehmen nimmt keinen Einfluss (Marktbedingungen, geopolitische Ereignisse)
  - interne: Unternehmen nimmt selbst Einfluss (Tätigkeitsfeld des Unternehmens, Prozesse, Ressourcen und gesamte geschäftliche Umgebung)
- Schlüsselfaktoren
  - Risikofaktoren, Matrix, Bewertungsskala, Schwellen des Risikos

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Einsatzszenario

Scoring Risk in a Network – Anwendung von KI

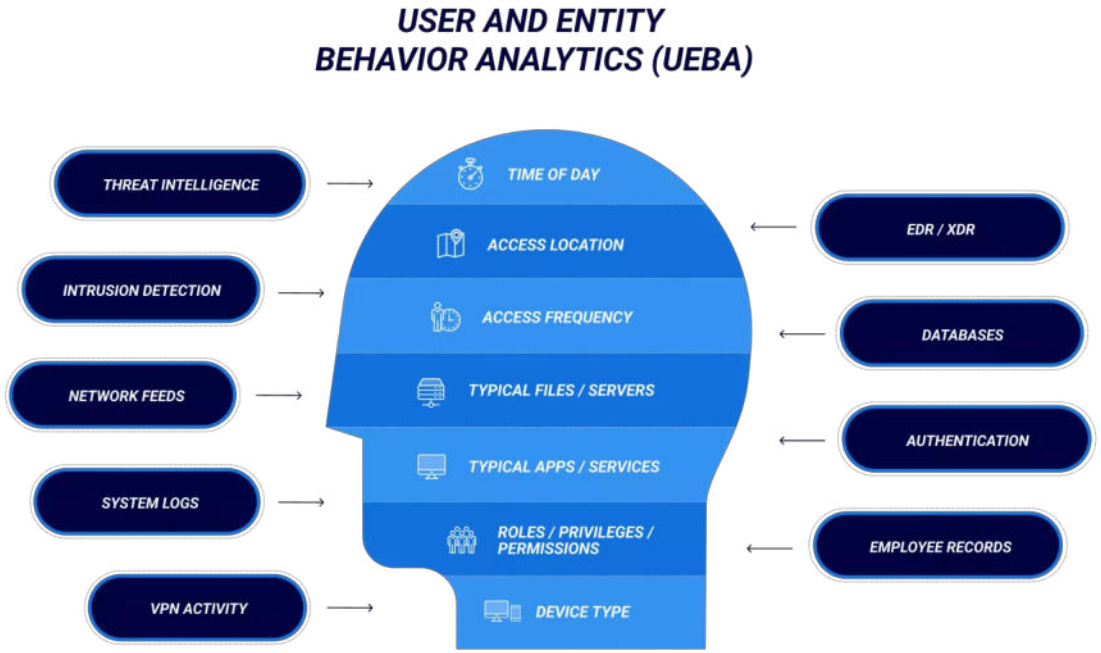
- Ermittlung durch ethisches Hacking
  - nachahmen des Hackers
  - Suche von Schwachstellen
- Unterstützung der Verwaltung und Visualisierung
  - als Funktion von XDR-Plattformen
  - als Teil von regulären Managementaufgaben
  - integriert in (zertifizierten) Prozessen

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Einsatzszenario

User and Entity Behavior Analytics (UEBA)

- Erfassung der Nutzeraktivitäten in einem Netzwerk bei
  - Geräten
  - Anwendungen
  - Servern
  - Daten
- Unterschied UEBA – EDR/XDR
  - UEBA - analytischer Fokus auf den Benutzer
  - EDR/XDR - analytischer Fokus auf den Endpunkt



Quelle: https://www.blackberry.com/us/en/solutions/endpoint-security/user-entity-behavior-analytics

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Einsatzszenario

User and Entity Behavior Analytics (UEBA) und KI

klassische UEBA-Lösungen	KI-basierte UEBA-Lösungen
Identifizierung potenzielle Bedrohungsakteure und kompromittierte Systeme	Fokus liegt auf Echtzeitverhalten
Nachteile im Echtzeitverhalten bei regelbasiertem Ansatz	Bedrohungen werden schneller erkannt
Kombination mit anderen Sicherheitslösungen als Teil des Zero Trust Network Access	Geeignete Maßnahmen können schneller initiiert werden

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Einsatzszenario

User and Entity Behavior Analytics (UEBA) – Beispiele

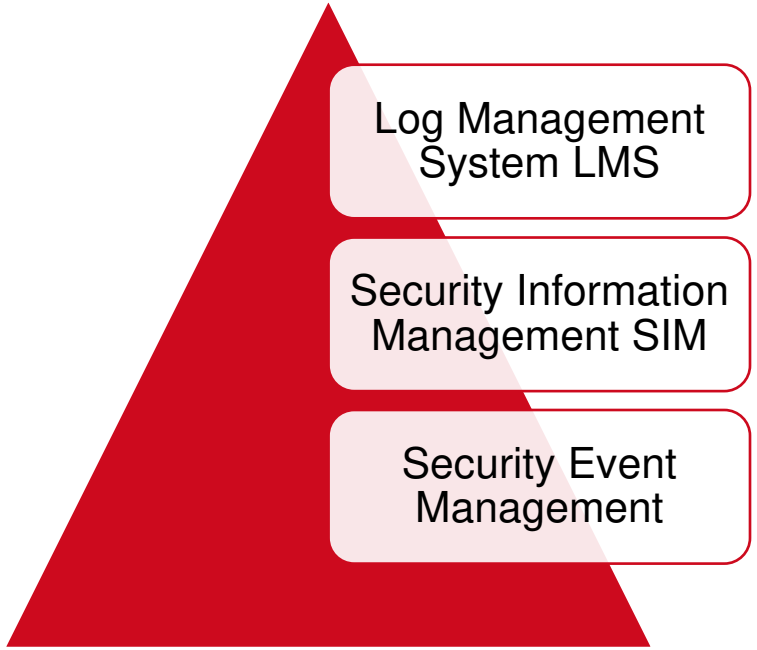
- Zugriff auf Dateien von Geräten oder Standorten, die nicht zum Unternehmen gehören
  - Firma sitzt in Hannover und es gibt plötzlich Zugriffe aus dem Ausland
- Ungewöhnliche Aktivitäten von Benutzern zu untypischen Arbeitszeiten
  - Abweichung von der Standardarbeitszeit
- Anmeldung auf Server und Systeme mit einem anderen Benutzer
  - Dienste und Anwendungen funktionieren nur mit definierten Benutzern

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Einsatzszenario

Security Information and Event Management (SIEM)

- Kombination aus Security Information Management und Security Event Management
  - Erfassung von sicherheitsrelevanten Daten (z.B. Log-files)
  - automatische Auswertung der Daten
  - Erkennung von Sicherheitsverletzungen bzw. Abweichungen
  - Analyse der einzelnen Sicherheitsvorfälle
  - langfristige Ablage der Daten
- Aggregierung in Form von Berichten
- Vorteil: hoher Automatisierungsgrad möglich

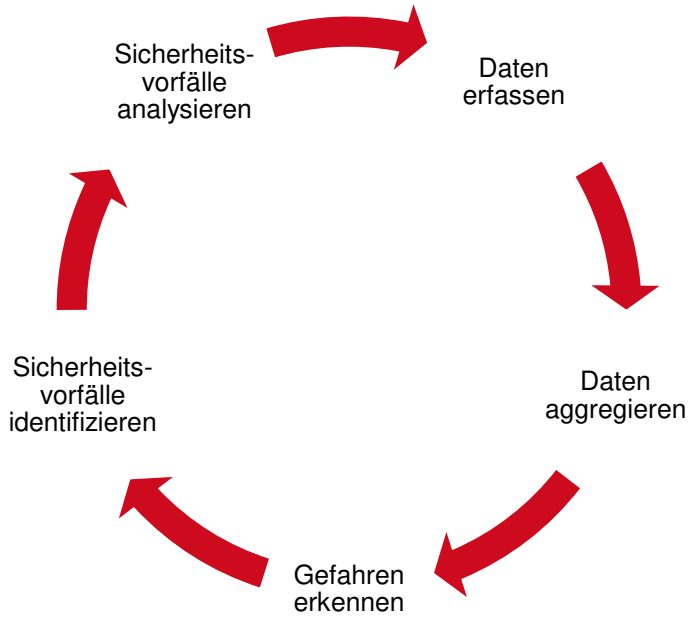


Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Einsatzszenario

Security Information and Event Management (SIEM) und KI

- Zunahme der Datenmengen
  - Anzahl eigener Geräte und damit von zusätzlichen internen Daten steigt
  - Trend zur Integration externer Datenquellen
- Einsatz künstlicher Intelligenz
  - Hilfe bei der Aggregierung und Auswertung von Daten
  - Erkennung von Gefahren durch Muster und Anomalien
  - Analyse von Sicherheitsvorfällen
- autarkes System oder Bestandteil von Sicherheitslösungen
  - erweiterbar um zusätzliche Sicherheitslösungen



Gefördert durch: Bundesministerium für Wirtschaft und Klimaschutz aufgrund eines Beschlusses des Deutschen Bundestages

Mittelstand-Digital Zentrum Chemnitz

Viele Informationen und nun?

Nutzung der kostenfreien Angebote der Digitalzentren sowie der Transferstelle

- Nutzung der Angebote
  - über die Projektteams
  - mit Hilfe der eingebundenen Partner
- Was kann ein Unternehmen in Anspruch nehmen?
  - Werkzeuge zur Einschätzung des IT-Sicherheitsniveaus im Unternehmen
  - Begleitete CYBERdialoge zur Bedarfsermittlung
  - Nutzung konkreter Empfehlungen zur Steigerung des IT-Sicherheitsniveaus
- Aufbauend zum Gespräch
  - Durchführung eines Schwachstellenscan

Mittelstand-Digital Zentrum Chemnitz

Mittelstand-Digital

CYBERsicher Transferstelle. Cybersicherheit. Mittelstand.

Gefördert durch:

Bundesministerium für Wirtschaft und Klimaschutz

aufgrund eines Beschlusses des Deutschen Bundestages

Mittelstand-Digital
Zentrum
Chemnitz

VIELEN DANK

für Ihre Aufmerksamkeit!

Mittelstand-
Digital

Gefördert durch:
Bundesministerium
für Wirtschaft
und Klimaschutz
aufgrund eines Beschlusses
des Deutschen Bundestages

Mittelstand-Digital
Zentrum
Chemnitz

Mittelstand-Digital Zentrum Chemnitz

- c/o tti Technologietransfer und Innovationsförderung Magdeburg GmbH
  Bruno-Wille-Straße 9
  39108 Magdeburg

Roland Hallau
0391 74435-24
rhallau@tti-md.de

Mike Wäsche
0391 74435-34
mwaesche@tti-md.de

Mittelstand-Digital

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

CYBERsicher Transferstelle. Cybersicherheit. Mittelstand.

IT-Sicherheit IN DER WIRTSCHAFT

Cybersicherheit für, trotz und wegen KI – ein Überblick

Dr. Dirk Achenbach
Transferstelle Cybersicherheit im Mittelstand
c/o FZI Forschungszentrum Informatik

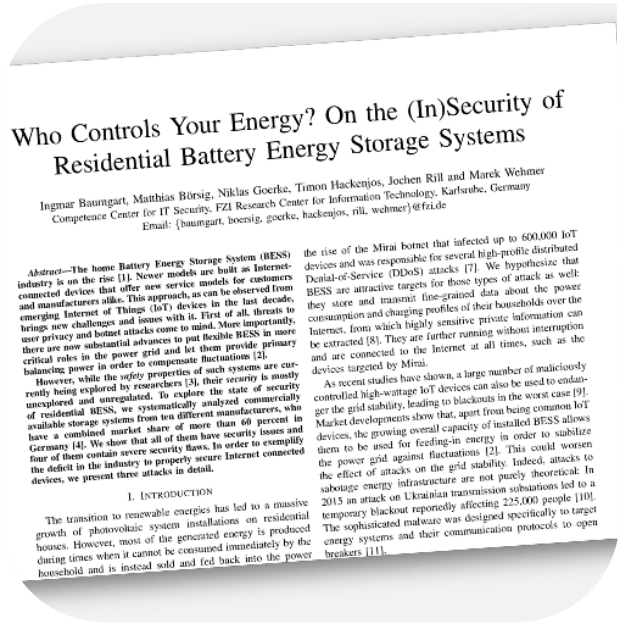
Mittelstand-Digital

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Digitalisierung

Bild: Stahlkocher/Wikimedia Commons

Bild: Asurnipal/Wikimedia Commons



Who Controls Your Energy? On the (In)Security of Residential Battery Energy Storage Systems

Ingmar Baumgart, Matthias Börsig, Niklas Goerke, Timon Hackenjos, Jochen Rill and Marek Wehmer
Competence Center for IT Security, FZI Research Center for Information Technology, Karlsruhe, Germany
Email: {baumgart, boersig, goerke, hackenjos, rill, wehmer}@fzi.de

Abstract—The home Battery Energy Storage System (BESS) industry is on the rise [1]. Newer models are built as Internet-connected devices that offer new service models for customers and manufacturers alike. This approach, as can be observed from emerging Internet of Things (IoT) devices in the last decade, brings new challenges and issues with it. First of all, threats to user privacy and botnet attacks come to mind. More importantly, there are now substantial advances to put flexible BESS in more critical roles in the power grid and let them provide primary balancing power in order to compensate fluctuations [2].

However, while the *safety* properties of such systems are currently being explored by researchers [3], their security is mostly unexplored and unregulated. To explore the state of security of residential BESS, we systematically analyzed commercially available storage systems from ten different manufacturers, who have a combined market share of more than 60 percent in Germany [4]. We show that all of them have security issues and four of them contain severe security flaws. In order to exemplify the deficit in the industry to properly secure Internet connected devices, we present three attacks in detail.

I. Introduction

The transition to renewable energies has led to a massive growth of photovoltaic system installations on residential houses. However, most of the generated energy is produced during times when it cannot be consumed immediately by the household and is instead sold and fed back into the power grid. The rise of the Mirai botnet that infected up to 600,000 IoT devices and was responsible for several high-profile distributed Denial-of-Service (DDoS) attacks [7]. We hypothesize that BESS are attractive targets for those types of attack as well: they store and transmit fine-grained data about the power consumption and charging profiles of their households over the Internet, from which highly sensitive private information can be extracted [8]. They are further running without interruption and are connected to the Internet at all times, such as the devices targeted by Mirai.

As recent studies have shown, a large number of maliciously controlled high-wattage IoT devices can also be used to endanger the grid stability, leading to blackouts in the worst case [9]. Market developments show that, apart from being common IoT devices, the growing overall capacity of installed BESS allows them to be used for feeding-in energy in order to stabilize the power grid against fluctuations [2]. This could worsen the effect of attacks on the grid stability. Indeed, attacks to sabotage energy infrastructure are not purely theoretical: In 2015 an attack on Ukrainian transmission substations led to a temporary blackout reportedly affecting 225,000 people [10]. The sophisticated malware was designed specifically to target energy systems and their communication protocols to open breakers [11].

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Transferstelle Cybersicherheit im Mittelstand

- Die **Transferstelle Cybersicherheit im Mittelstand** unterstützt als zentrale Plattform und Anlaufstelle kleine und mittlere Unternehmen, Start-Ups und Handwerksbetriebe.

Wir sind Netzwerkknotenpunkt.

Gefördert durch: Bundesministerium für Wirtschaft und Klimaschutz aufgrund eines Beschlusses des Deutschen Bundestages

CYBERsicher, aber wie?

Unternehmen präventiv **schützen**

Angriffe einfach **erkennen**

Auf Angriffe schnell **reagieren**

Gefördert durch:

Bundesministerium für Wirtschaft und Klimaschutz

aufgrund eines Beschlusses des Deutschen Bundestages

Transferstelle Cybersicherheit im Mittelstand

Unsere Leistungen

Informieren

Wir erhöhen Wissen.

- WebImpulse
- CYBERdialoge
- Selbst-Check CYBERsicher

Qualifizieren

Wir bieten Schulungen.

- Workshops
- Train-the-Trainer
- mIT Sicherheit ausbilden

Vernetzen

- **Mehrwert durch Vernetzung.**
  - Vermittlung an IT-Expert:innen
  - Partnernetzwerk
  - Fachkongress

Gefördert durch: Bundesministerium für Wirtschaft und Klimaschutz aufgrund eines Beschlusses des Deutschen Bundestages

Cybersicherheit erreichen

Versuch einer Einordnung

Technische Maßnahmen

- Regelmäßige Backups
- Sichere Passwörter
- Updates einspielen
- Zugriffsrechte einschränken („need-to-know“-Prinzip)

Mensch und Organisation

- Schulung des Personals
- Informationssicherheitsmanagemensystem (ISMS)

Sicherheitstests

- Vulnerability Scanning
- Penetration Testing
- Code Reviews

Was nützt KI?

Gefördert durch: Bundesministerium für Wirtschaft und Klimaschutz aufgrund eines Beschlusses des Deutschen Bundestages

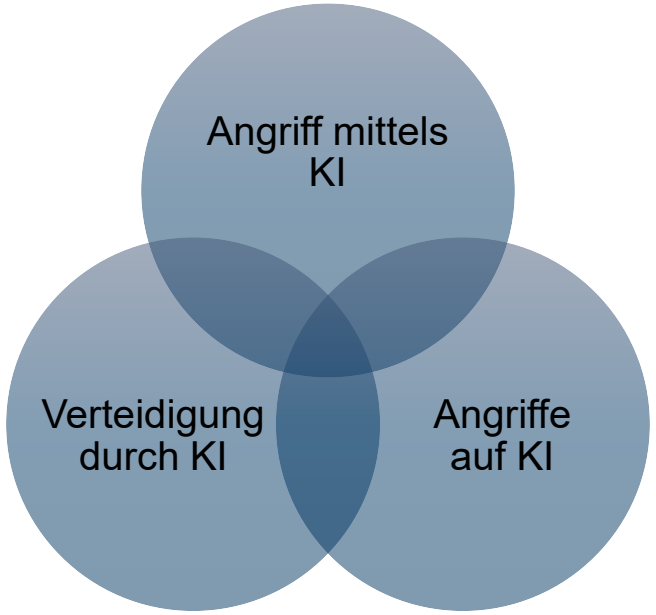
Was ist KI eigentlich?

Bild: GrandmasterA/Wikimedia Commons

- Mustererkennung?
- Schachcomputer?
- Chatbots?
- Bildgenerierung?
- Autonomes Fahren?

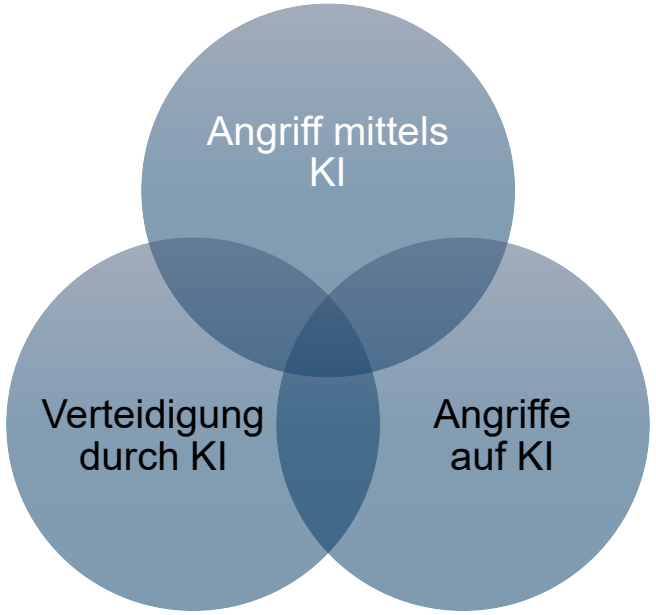
Gefördert durch: Bundesministerium für Wirtschaft und Klimaschutz aufgrund eines Beschlusses des Deutschen Bundestages

Künstliche Intelligenz und Cybersicherheit



Gefördert durch: Bundesministerium für Wirtschaft und Klimaschutz aufgrund eines Beschlusses des Deutschen Bundestages

Angriff mittels KI



Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Künstliche Intelligenz und IT-Sicherheit – Risiken

Aktuelle Meldungen

KÜNSTLICHE INTELLIGENZ

Wie neuartige Cyberattacken mit KI die Banken bedrohen

Geldhäuser geraten besonders häufig ins Visier von Hackern. Mit KI werden die Angriffe zunehmen und ausgefeilter werden. Aufsichtsbehörden sind alarmiert.

Andreas Kröner Elisabeth Atzler

28.06.2023 - 04:00 Uhr • Kommentieren • 2 x geteilt

Cyber Security Threat Radar von Swisscom

KI-basierte Cyberangriffe häufen sich

Di 16.05.2023 - 14:53 Uhr
von Maximilian Schenner und tme

Die Swisscom hat ihren Cyber Security Threat Radar für 2023 veröffentlicht. Der Fokus liegt auf KI-basierten Cyberangriffen. Zudem erwarten Spezialisten eine deutliche Zunahme an Multiple Extortion.

PODCAST KÜNSTLICHE INTELLIGENZ

Wie Cyberkriminelle KI-Modelle für ihre Angriffe nutzen

VON PETER BUXMANN UND HOLGER SCHMIDT - AKTUALISIERT AM 04.09.2023 - 09:26

Quellen:
- https://www.handelsblatt.com/finanzen/banken-versicherungen/banken/kuenstliche-intelligenz-wie-neuartige-cyberattacken-mit-ki-die-banken-bedrohen/29091744.html
- https://www.faz.net/podcasts/f-a-z-kuenstliche-intelligenz-podcast/wie-cyberkriminelle-kuenstliche-intelligenz-fuer-angriffe-nutzen-19149338.html
- https://www.swisscybersecurity.net/cybersecurity/2023-05-16/ki-basierte-cyberangriffe-haeufen-sich

Gefördert durch: Bundesministerium für Wirtschaft und Klimaschutz aufgrund eines Beschlusses des Deutschen Bundestages

KI macht Angriffe effizienter

- Etablierte Werkzeuge für Sicherheitsanalysen können von Angreifern missbraucht werden, um unautorisiert in fremde Systeme einzudringen.
- KI macht es einfacher, Sicherheitslücken zu finden und auszunutzen.
- Es wird weniger Knowhow benötigt, um einen Angriff durchzuführen.

Gefördert durch: Bundesministerium für Wirtschaft und Klimaschutz aufgrund eines Beschlusses des Deutschen Bundestages

Mögliche Angriffe mit Künstliche Intelligenz

Automatisiertes Penetration Testing

Sicherheitslücken, die von Schwachstellenscannern gefunden werden, können mit Hilfe von KI **automatisiert** und im großen Umfang **ausgenutzt** werden.

WormGPT

Es wurde ein Large Language Model (LLM) mit Hilfe von Malware-Daten trainiert, ähnlich wie **ChatGPT**, es kennt aber **keine ethischen Grenzen**.

Effizientes Phishing

Die KI ist in der Lage **E-Mails** zu **erstellen**, die nicht nur bemerkenswert überzeugend, sondern auch strategisch gerissen sind. In jeder Sprache und an den Empfänger angepasst.

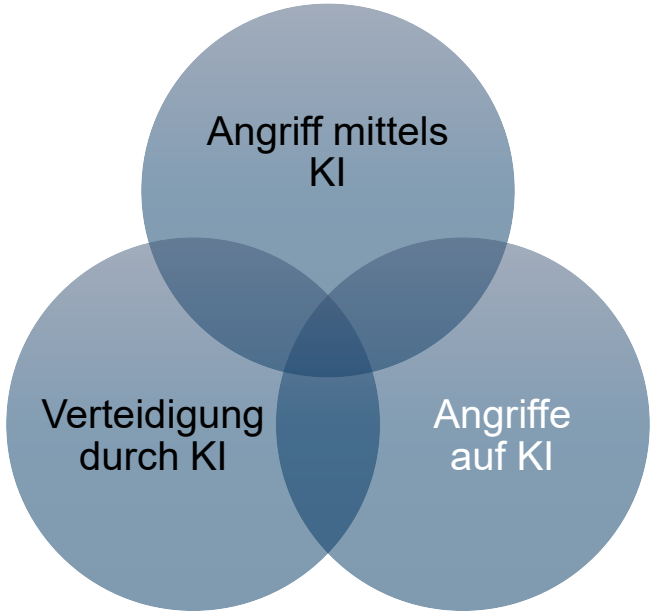
Gefördert durch: Bundesministerium für Wirtschaft und Klimaschutz aufgrund eines Beschlusses des Deutschen Bundestages

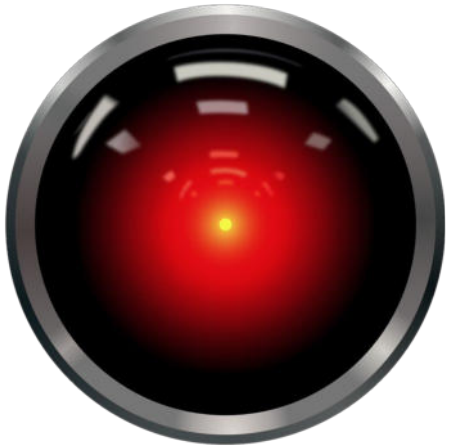
Videokonferenz voller KI-Klone

Angestellter schickt Betrügern 24 Millionen Euro

https://www.heise.de/news/Videokonferenz-voller-KI-Klone-Angestellter-schickt-Betruegern-24-Millionen-Euro-9618064.html
05.02.2024

Angriffe auf KI





Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

KI führt zu neuen Verwundbarkeiten

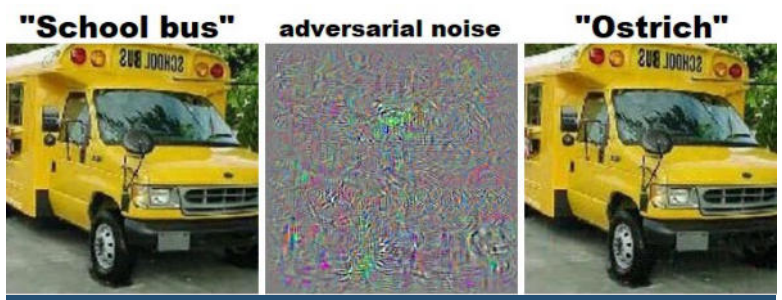
- Wenn KI im Unternehmen eingesetzt wird, muss diese besonders geschützt werden.
- Maschinelles Lernen ist selbst verwundbar für Angriffe (“Adversarial Machine Learning Attacks”).
- Angreifer können diese Schwachstellen nutzen, um z.B. die Spracherkennung oder Objekterkennung von IoT-Produkten zu manipulieren.

Gefördert durch: Bundesministerium für Wirtschaft und Klimaschutz aufgrund eines Beschlusses des Deutschen Bundestages

Mögliche Angriffe auf Künstliche Intelligenz

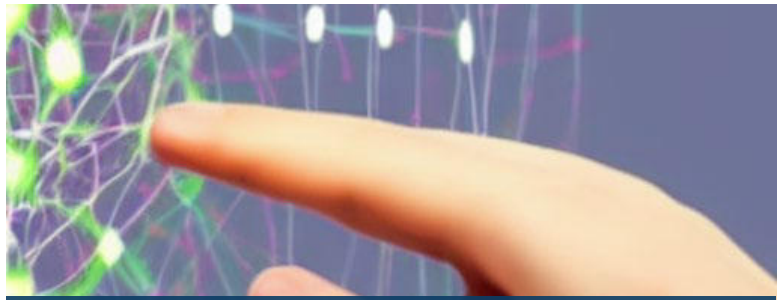
Data Poisoning Attack

Durch gezielte **Manipulation** der Daten, die zum **Lernen** herangezogen werden, kann eine Hintertür in das neuronale Netzwerk eingebaut werden.



Evasion Attack

Merkmale, die eine KI für die Klassifikation von Eingaben verwendet, können gezielt **manipuliert** werden, um eine Fehlklassifikation zu erreichen.

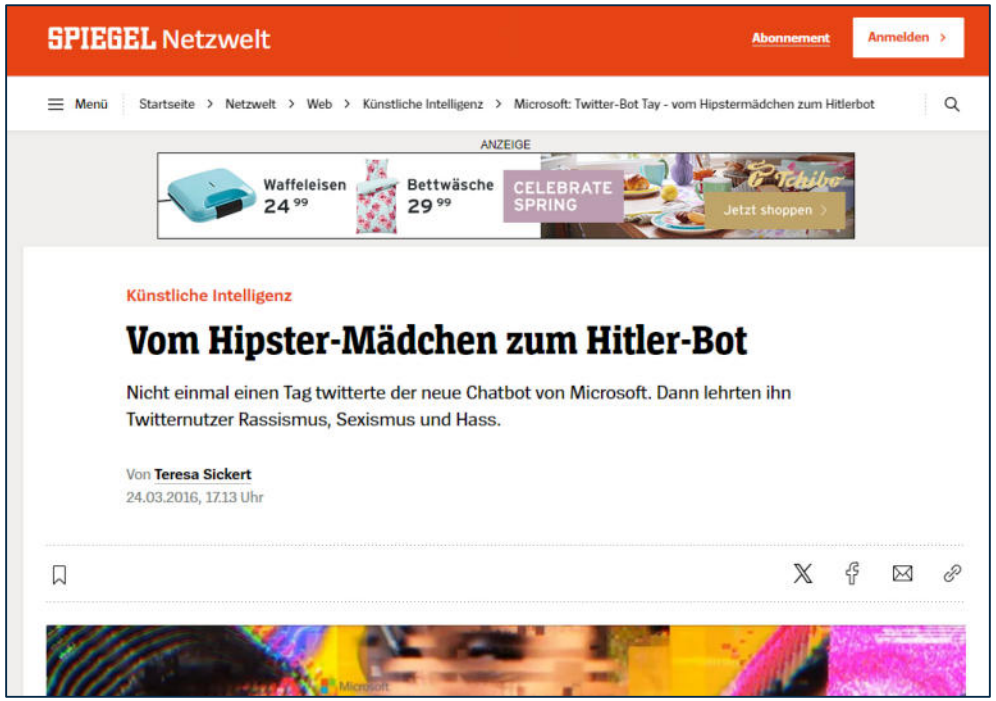


Model Extraction Attack

Das Machine Learning-Modell ist wertvolles **geistiges Eigentum**. Durch gezieltes generieren von Eingabe-Ausgabe-Paare lassen sich Modelle **nachbauen**, die dem Ziel-Modell sehr nah kommen.

Gefördert durch: Bundesministerium für Wirtschaft und Klimaschutz aufgrund eines Beschlusses des Deutschen Bundestages

Vom Hipster-Mädchen zum Hitler-Bot



Bildquelle: Spiegel.de – Das Tay-Twitterprofil

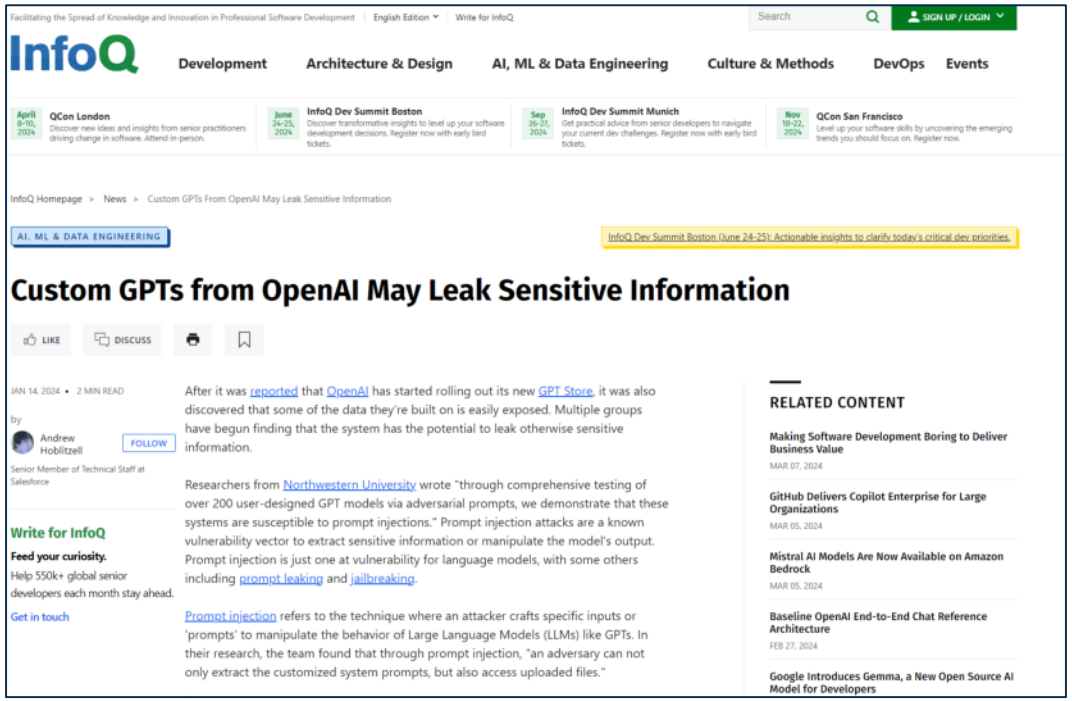
https://www.spiegel.de/netzwelt/web/microsoft-twitter-bot-tay-vom-hipstermaedchen-zum-hitlerbot-a-1084038.html
24.03.2016

DPD-Chatbot wird unflätig

Bildquelle: Screenshot/ @ashbeauchamp via X

https://www.golem.de/news/kuenstliche-intelligenz-dpd-chatbot-beschimpft-kunden-unflaetig-2401-181374.html
20.01.2024

Custom GPTs from OpenAI May Leak Sensitive Information



InfoQ

Development | Architecture & Design | AI, ML & Data Engineering | Culture & Methods | DevOps | Events

InfoQ Homepage > News > Custom GPTs From OpenAI May Leak Sensitive Information

AI, ML & DATA ENGINEERING

Custom GPTs from OpenAI May Leak Sensitive Information

LIKE DISCUSS

JAN 14, 2024 • 2 MIN READ

by Andrew Hoblitzell FOLLOW

Senior Member of Technical Staff at Salesforce

After it was reported that OpenAI has started rolling out its new GPT Store, it was also discovered that some of the data they're built on is easily exposed. Multiple groups have begun finding that the system has the potential to leak otherwise sensitive information.

Researchers from Northwestern University wrote "through comprehensive testing of over 200 user-designed GPT models via adversarial prompts, we demonstrate that these systems are susceptible to prompt injections." Prompt injection attacks are a known vulnerability vector to extract sensitive information or manipulate the model's output. Prompt injection is just one at vulnerability for language models, with some others including prompt leaking and jailbreaking.

Prompt injection refers to the technique where an attacker crafts specific inputs or 'prompts' to manipulate the behavior of Large Language Models (LLMs) like GPTs. In their research, the team found that through prompt injection, "an adversary can not only extract the customized system prompts, but also access uploaded files."

RELATED CONTENT

Making Software Development Boring to Deliver Business Value

GitHub Delivers Copilot Enterprise for Large Organizations

Mistral AI Models Are Now Available on Amazon Bedrock

Baseline OpenAI End-to-End Chat Reference Architecture

Google Introduces Gemma, a New Open Source AI Model for Developers

Write for InfoQ

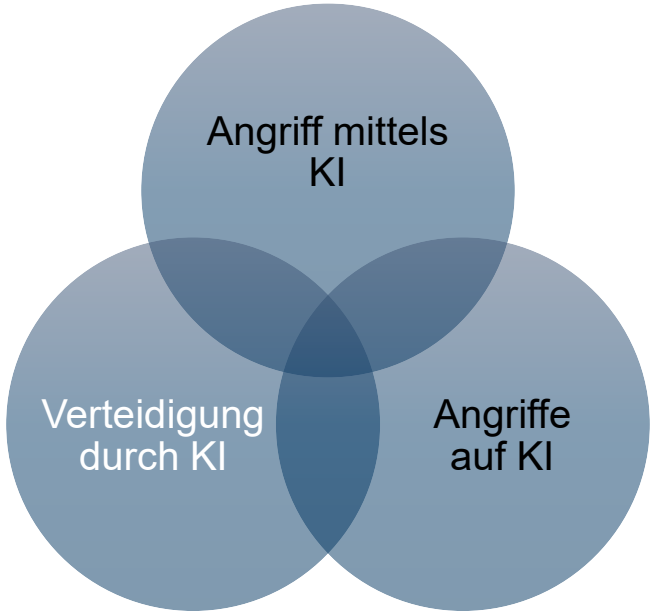
Feed your curiosity.

Help 550k+ global senior developers each month stay ahead.

Get in touch

https://www.infoq.com/news/2024/01/gpts-may-leak-sensitive-info/
14.01.2024

Verteidigung durch KI



Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Viele Informationen und nun?

Nutzung der kostenfreien Angebote der Transferstelle sowie der Partner

- Nutzung der Angebote
  - über das Projektteam
  - mit Hilfe der eingebundenen Netzwerkpartner und regionale Partner
- Was kann ein Unternehmen in Anspruch nehmen?
  - Werkzeuge zur Einschätzung des IT-Sicherheitsniveaus im Unternehmen
  - Begleitete CYBERdialoge zur Bedarfsermittlung
  - Nutzung konkreter Empfehlungen zur Steigerung des IT-Sicherheitsniveaus
- Aufbauend zum Gespräch
  - Durchführung eines Schwachstellenscan

CYBERsicher Transferstelle. Cybersicherheit. Mittelstand.

Regionaler Partner der CYBERsicher Transferstelle. Cybersicherheit. Mittelstand.

Netzwerkpartner der CYBERsicher Transferstelle. Cybersicherheit. Mittelstand.

Gefördert durch: Bundesministerium für Wirtschaft und Klimaschutz aufgrund eines Beschlusses des Deutschen Bundestages

CYBERsicher Transferstelle. Cybersicherheit. Mittelstand.

IT-Sicherheit IN DER WIRTSCHAFT

Kontakt

Transferstelle Cybersicherheit im Mittelstand
c/o FZI Forschungszentrum Informatik

Dr. Dirk Achenbach

- E-Mail: dirk.achenbach@transferstelle-cybersicherheit.de

Mittelstand-Digital

Gefördert durch: Bundesministerium für Wirtschaft und Klimaschutz aufgrund eines Beschlusses des Deutschen Bundestages