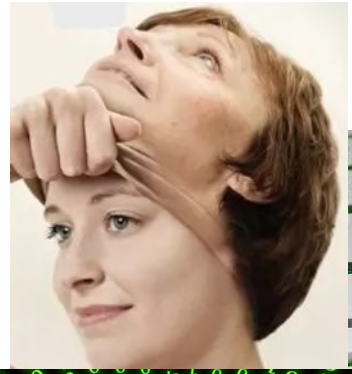
IT-Sicherheit, bes. KI

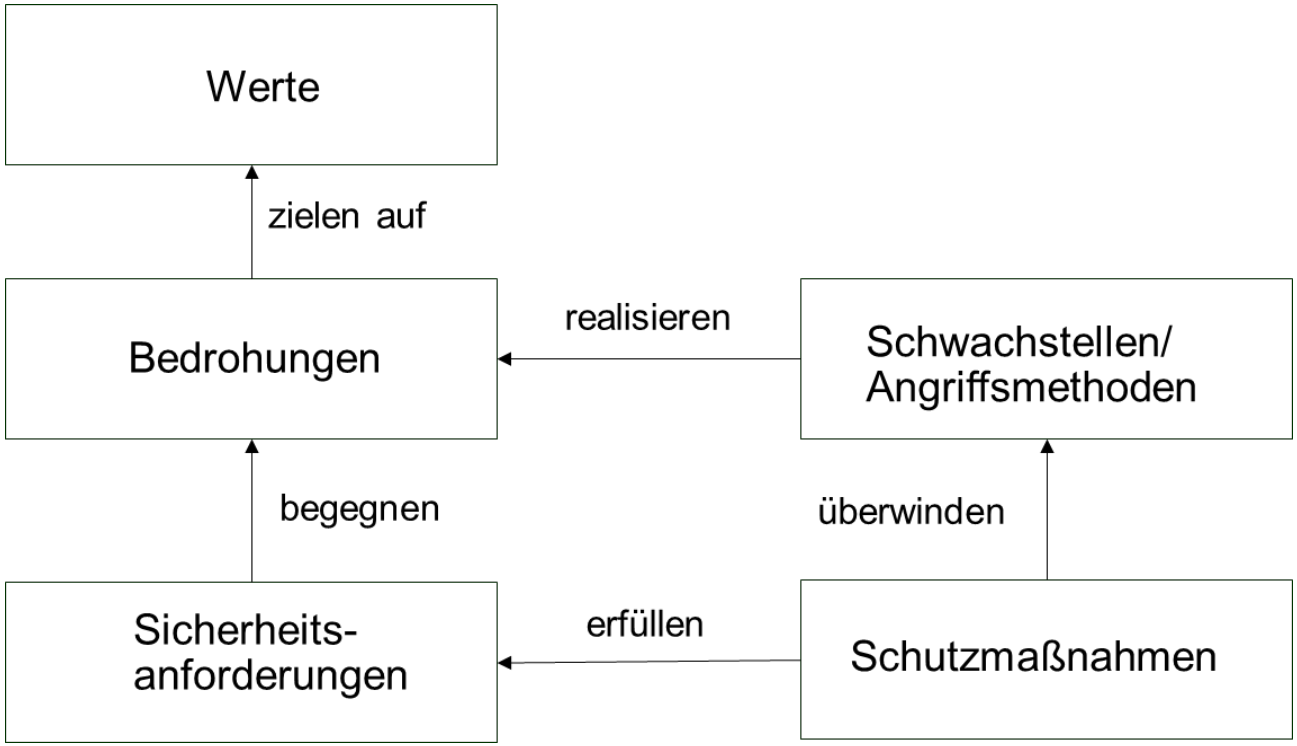
Stand der Technik



Saarbrücken
EDV-Tag Juris
11. September 2024

Prof. Dr. Rüdiger Grimm
Fellow GI
Universität Koblenz
Fraunhofer SIT Darmstadt

I. Begriffliche Einordnung



II. Bedrohungen - Sicherheitsanforderungen

1. Missbrauch personenbezogener Daten – Privatheit/Datenschutz
2. Lauschen und Spionage – Vertraulichkeit
3. Fälschen – Integrität
4. Identitätsdiebstahl – Korrekte Herkunft (Originalität)
5. Dienstverweigerung (Denial of Service) – Verfügbarkeit
6. Ableugnen – Nicht-Abstreitbarkeit
7. Eindringen und Hacking – Zugriffskontrolle

II a. Moderne Bedrohungen

8. Kriminalität
   - Angriffe mit dem Ziel der illegalen Bereicherung
   - zB. Lahmlegen – Erpressen (Unis Gießen 2019, Essen 2022)
   - Vorbereitung krimineller Aktivitäten
   - Verwischen von Spuren, z.B. Anonymität, Darknet
9. Cyberwar
   - Gegen kritische Infrastrukturen
   - Desinformation
   - Spionage
10. ***KI – das unbekannte Wesen***

s.u. #10 ff

III. Besondere Angriffstechniken

1. Kryptoanalyse
2. Viren, Würmer, Trojanische Pferde
3. Bot-Netze: Kapern von PCs
4. SPAM, Phishing, Pharming: Stehlen von Zugangsdaten
5. DDoS – Distributed Denial of Service
6. Tracking/Tracing
7. Unautorisierte Ausforschung und Big Data
8. Medienfälschung
9. ***Maschinelles Lernen zur Unterstützung von Angriffen*** s.u. #13

IV. Grundlegende Schutzmaßnahmen

1. Rechtliche, technische und organisatorische Maßnahmen ⇦
2. Kryptografie
3. Digitale Signatur
4. Post-Quantum-Verschlüsselung ⇦
5. Smartcards und TPM
6. Digitale Wasserzeichen
7. Authentifizierung und Zugriffskontrolle
8. ***Maschinelles Lernen zur Abwehr von Angriffen*** ⇦ s.u. #15
9. Abwehr, Aufdeckung und Prävention
10. Security by Design und Privacy by Design

Zu „IV.1 Rechtliche + technische + organisatorische Maßnahmen“

- Beispiel Persönlichkeitsschutz – Löschfristen
  - Datenschutz GVO
  - Löschverfahren
  - Datenschutzbeauftragte, Schulungen usw.
- Beispiel Urheberschutz
  - UrhG, bes. §106 Strafbewehrung; Orga-RL (z.B. DFN)
  - Wasserzeichen, Ähnlichkeitserkennung, Plagiat-SW
  - Medienbeauftragte, Aufklärung usw.
- Beispiel Cyberverteidigung
  - Meldepflichten NIS-2 (EU-RL seit 2022)
  - Erkennung von Angriffen
  - Sicherheitsbeauftragte, Schulungen usw.

Zu IV.4 „Post-Quantum-Verschlüsselung“

- Quantencomputer (QC)
  - unentschiedene Quantenzustände für hochparalleles Rechnen
  - Brechen von Kryptoverfahren durch Brute-Force-Rechnung
  - In 10-20 Jahren einsatzbereit
- QC-Resistenz für langlebige Produkte und Verfahren
  - Medizin
  - industrielle Produktionssteuerung
  - Entwicklung von Autos, Flugzeugen und Raumfahrt

… 4. Post-Quantum-Verschlüsselung

- NIST-Kandidaten für QC-Resistenz
  - selbstkorrigierende Code-Fehler-Verfahren
  - Hashbäume
  - Gleichungssysteme mit multivariaten Polynomen
  - mathematische Gitter
  - sogenannte Isogenien supersingulärer elliptischer Kurven

  - CRYSTALS-KYBER und CRYSTALS-Dilithium
  - ***To do:*** Verbesserung, Inventarisierung, normierte Schnittstellen

V. KI – Maschinelles Lernen

- Computerprogramme entwickeln Algorithmen und Modelle, die in großen Datenmengen Muster und Beziehungen erkennen und daraus zugehörige Vorhersagen oder Entscheidungen ableiten

- Nützliche Anwendungen

- Aber auch: Angriffe

Starke und schwache KI

- Starke KI
  - wie Menschen selbstständig Situationen verstehen, aus ihnen lernen und daraufhin in neuen Situationen Urteile fällen und Probleme lösen („Turing-Test“)
  - „…not yet a reality…“
- Schwache KI, Maschinelles Lernen
  - Data Mining: große Datenmengen analysieren und für diese Antworten auf konkrete Anwendungsfragen entwickeln
  - „überwachtes“ bzw. „unüberwachtes Lernen“

„Überwachtes“ bzw. „unüberwachtes Lernen“

- „Überwachtes Lernen“
  - Ziel: korrekte Aussagen über unbekannte Daten
    zB Bilder: „Das ist eine Katze“
  - Trainingsphase mit „wahr/falsch“-markierten Bilden
  - Klassifizierung von Daten, zB Bilder, Spam, kriminelle Kommunikation, Betrug
  - Prognosen zukünftiger Ereignisse, zB Aktienkurse, kriminelle Handlungen
- „Unüberwachtes Lernen“
  - Ziel: bisher unbekannte Muster in oder Beziehungen zwischen Daten finden
  - Nicht für Prognosen, sondern
  - Reduktion von Eigenschaftsdimensionen, zB Datenkompression
  - Gruppenbildung (Clustering), zB Anomalien (Muster auch ohne Trainingsphasen), zB Kundensegmentierung, zB Kaufempfehlungen

Zu „III.9 Angriffe mit KI – Maschinelles Lernen“

- Dynamisches Fälschen von Videomaterial („Deep Fake“)
  https://www.sit.fraunhofer.de/fileadmin/videos/Anwendertag_2024_01_final.mp4
- Einstreuen von Propaganda in politische Internetforen
  (Wahlkampf)
- Suche nach Schwachstellen im Netz
- Design von Authentifizierungsmaterial
  (Bilder, Texte, Passwörter)
- Irreführung „guter“ Lernverfahren
  Verfälschung der Trainingsdaten oder der Zielfunktion

- **Irreführung „guter“ Lernverfahren**

Beispiel **Adversarial Examples:**

Fälschung von Trainingsdaten,
die gezielt verändert werden,
um Fehlentscheidung des Modells zu provozieren

zB durch Löschen von Erkennungsmerkmalen
zB durch Einbringen falscher Erkennungsmerkmale

„Ich bin eine Katze“

„Leckere Birnen“

Zu „IV.8 Abwehr mit KI – Maschinelles Lernen“

- Erkennung von Kommunikationsanomalien im Netz („Intrusion Detection“)
- KI-Klassifikation: Filtern von
  - unangemessene Bilder aus massenweisem Bildmaterial
  - kriminelle Kommunikation in Chatforen
  - Betrugsmaschen oder ungewöhnliches Kaufverhalten im E-Commerce
  - unstimmiges Authentifizierungsmaterial in lernbaren Mustern oder Datenrelationen
- KI-Verteidigung von KI-Angriffen

… IV.8 … KI-Abwehr von KI-Angriffen

- GAN – Generative Adversarial Networks
  - **Zwei** konkurrierende Künstliche Neuronale Netze:
  - **Generator** als Angreifer generiert künstliche Daten
  - **Diskriminator** als Verteidiger erkennt Daten als echt oder künstlich
- zB. Authentifizierungsmaterial
  - Eindringling lernt täuschen
  - Diskriminator lernt erkennen
- Generator = „böse“? Nicht unbedingt:
  - Hochqualitative künstlich erzeugte Daten sind wertvolle Trainingsdaten für „gute“ KI-Lernverfahren
  - zB Ziel, mithilfe von GAN Privatheit-schützende Trainingsdaten herzustellen

Fazit

- KI ist nicht gut oder böse
- Nutzung von KI erfordert Transparenz
- Herrschaftsanspruch: Wer trainiert die Systeme?
- Welche Sprachkultur dominiert ChatGPT?
- KI erfordert aktive Gestaltung
- Jugend muss KI kennen und einschätzen lernen

Literaturverweis:

Hornung, G.; und Schallbruch, M. (Hrsg.): IT-Sicherheitsrecht, Praxishandbuch, Nomos, 2. Auflage 2024, ca. 1000 Seiten, gebunden ISBN 978-3-7560-0496-6.

Darin §2: IT-Sicherheit aus technischer Sicht (Grimm/Waidner)