IFA
Institut für Arbeitsschutz der
Deutschen Gesetzlichen Unfallversicherung

Vertrauenswürdige künstliche Intelligenz

AUVA Webinar Industrie 4.0

Handout

12.04.2021

Dr. André Steimers

Institut für Arbeitsschutz der DGUV

Beispiele einiger KI Fehler

- 07/2016 Wachroboter verletzt Kind in Kaufhaus
- 11/2016 Roboter Xiao-Pang verletzt Messebesucher
- 05/2017 Auffahrunfall Tesla Modell S Feuerwehrfahrzeug
- 01/2018 Auffahrunfall Tesla Modell S Feuerwehrfahrzeug
- 05/2018 Auffahrunfall Tesla Modell S Polizeifahrzeug
- 01/2019 Kollission mit Gegenverkehr Tesla Modell 3
- 08/2019 Auffahrunfall Tesla Modell S Abschleppwagen
- 12/2020 Fehlfunktion eines Serviceroboters im Kaufhaus

Beispiele einiger KI Fehler

- 01/2016 China, Tesla Modell S,1 Fahrer†
- 05/2016 Florida, Tesla Modell S, 1 Fahrer†
- 03/2018 Arizona, automatisiertes Uber Taxi, 1 Fußgängerin†
- 03/2018 Kalifornien, Tesla Modell X, 1 Fahrer†
- 04/2018 Japan, Tesla Modell X, 1 Fußgänger†
- 03/2019 Florida, Tesla Modell 3, 1 Fahrer†
- 04/2019 Florida, Tesla Modell S, 1 Fußgänger†
- 12/2020 Kalifornien, Tesla Modell S, 2 Personen Honda Civic†
- 05/2020 Norwegen, Tesla Modell X, 1 Fußgänger†

Vertrauenswürdige KI

- High Level Expert Group for Artificial Intelligence:

  „Ethik Leitlinien für eine vertrauenswürdige KI“

3 Komponenten einer vertrauenswürdigen KI:

1. Sie sollte rechtmäßig sein und somit geltendes Recht und alle gesetzlichen Bestimmungen einhalten
2. Sie sollte ethisch sein und somit die Einhaltung ethischer Grundsätze und Werte garantieren
3. Sie sollte robust sein, und zwar sowohl in technischer als auch in sozialer Hinsicht, da KI-Systeme möglicherweise unbeabsichtigten Schaden verursachen, selbst wenn ihnen gute Absichten zugrunde liegen

Rechtliche Aspekte

- Bürgerliches Gesetzbuch
- Produkthaftungsgesetz
- Straßenverkehrsgesetz

Rechtliche Aspekte

Bürgerliches Gesetzbuch

- Paragraph 823 Schadensersatzpflicht

  „(1) Wer vorsätzlich oder fahrlässig das Leben, den Körper, die Gesundheit, die Freiheit, das Eigentum oder ein sonstiges Recht eines anderen widerrechtlich verletzt, ist dem anderen zum Ersatz des daraus entstehenden Schadens verpflichtet.“

- Verschuldensprinzip: wenn der Mensch, der ein KI System nutzt, es ordnungsgemäß gewartet und eingesetzt hat, haftet er grundsätzlich nicht
- Allgemeines Lebensrisiko
- Problem: es können Schäden entstehen für die keiner haftet

Rechtliche Aspekte

Produkthaftungsgesetz

- Paragraph 1 Haftung

„(1) Wird durch den Fehler eines Produkts jemand getötet, sein Körper oder seine Gesundheit verletzt oder eine Sache beschädigt, so ist der Hersteller des Produkts verpflichtet, dem Geschädigten den daraus entstehenden Schaden zu ersetzen.“

- Problem 1: kommt nur bei Körper- und Sachschäden zur Geltung
- Problem 2: bei KI-Systemen die der Vorhersage und Optimierung dienen ist meist unbekannt, ob das Ergebnis stimmt oder ein Fehler vorliegt
- Problem 3: die Komplexität der Aufgaben von KI-Systemen führt meist zu diffusen Systemspezifikationen

Rechtliche Aspekte

Straßenverkehrsgesetz

- Paragraph 7 Haftung des Halters

  „(1) Wird bei dem Betrieb eines Kraftfahrzeugs…. ein Mensch getötet, der Körper oder die Gesundheit eines Menschen verletzt oder eine Sache beschädigt, so ist der Halter verpflichtet, dem Verletzten den daraus entstehenden Schaden zu ersetzen.“

- Wenn ein selbstfahrendes Fahrzeug jemandem Schaden zufügt, kommt grundsätzlich die Kfz-Versicherung des Fahrzeughalters für den Schaden auf
- **Diskussion: Allgemeine Gefährdungshaftung für KI-Systeme**
  Wer ein KI-System einsetzt muss davon ausgehen, dass dabei Schäden entstehen können - und haftet deswegen unabhängig vom Verschulden (siehe Tierhalterhaftung)

Ethische Aspekte

1. Fairness
2. Privacy
3. Automatisierungsgrad und Kontrolle

1. Fairness

- **Rekrutierungs-Tool**
  diskriminiert Frauen

- **Historischer Bias**
  ML-Modell kann negative Korrelation lernen, da Männer in der Vergangenheit oft systematisch bevorzugt wurden

- **Gesichtserkennung**
  schlechtere Performance bei farbigen Menschen

- **Daten Bias**
  Unterrepräsentierte Gruppen in den Trainingsdaten führen zu höheren Fehlerraten dieser Gruppen im ML-Modell

2. Privacy

EU Datenschutz Grundverordnung 2016/679

Artikel 5 Paragraph 1	Persönliche Daten sollten …
Rechtmäßigkeit, Fairness, Transparenz	„… auf **rechtmäßige** Weise, nach **Treu und Glauben** und in einer für die betroffene Person **nachvollziehbaren Weise** verarbeitet werden…“
Zweckbindung	„… für **festgelegte**, **eindeutige** und **legitime Zwecke** erhoben werden …“
Datenminimierung	„… dem Zweck angemessen und erheblich sowie auf das **für die Zwecke** der Verarbeitung **notwendige Maß beschränkt** sein …“
Richtigkeit	„… sachlich **richtig** und … **auf dem neuesten Stand** …“
Speicherbegrenzung	„… in einer Form **gespeichert** werden, die die Identifizierung der betroffenen Personen **nur so lange** ermöglicht, **wie es** für die Zwecke, für die sie verarbeitet werden, **erforderlich ist**; …“
Integrität und Vertraulichkeit	„… in einer Weise verarbeitet werden, die eine **angemessene Sicherheit** der personenbezogenen Daten gewährleistet, …“

3. Automatisierungsgrad- und Kontrolle

0. Keine Automatisierung

- Bediener kontrolliert das System vollständig
- human control

1. Assistenzsysteme

- System unterstützt einen Bediener
- human-in-the-loop

2. Teilautomatisierung

- Teilfunktionen sind automatisiert aber Gesamtsystem bleibt unter externer Kontrolle
- human-in-the-loop

3. Bedingte Automatisierung

- Alle Funktionen sind automatisiert aber ein externer Eingriff muss jederzeit möglich sein
- human-on-the-loop

4. Hochautomatisierung

- Das System führt Teile seiner Aufgabe ohne externen Eingriff aus
- human-on-the-loop

5. Vollautomatisierung

- Das System ist in der Lage, seine gesamte Mission ohne externen Eingriff zu erfüllen

Sicherheitstechnische Aspekte

4. Grad der Transparenz und Erklärbarkeit
5. Zuverlässigkeit und Robustheit des Modells
6. Security
7. System-Hardware
8. Technologische Ausgereiftheit

4. Grad der Transparenz und Erklärbarkeit

1. Erklärbar:

   Das System liefert klare und kohärente Erklärungen.

2. Artikulierbar:

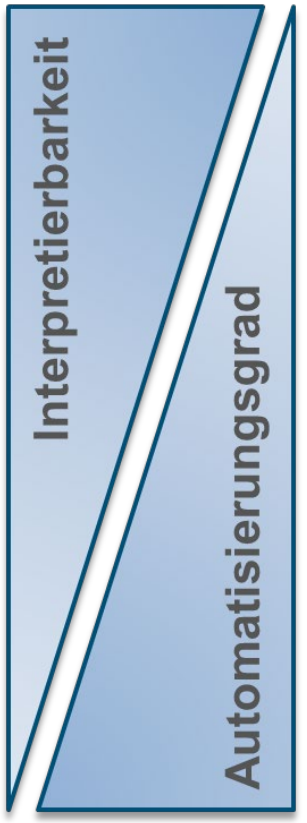
   Das System ist in der Lage die relevantesten Merkmale zu extrahieren und ihre Zusammenhänge und Wechselwirkungen grob darzustellen.

3. Nachvollziehbar:

   Das System ist nicht dazu in der Lage in Echtzeit eine Erklärung zum Systemverhalten zu liefern, diese sind aber zumindest im Nachhinein überprüfbar.

4. Black-Box:

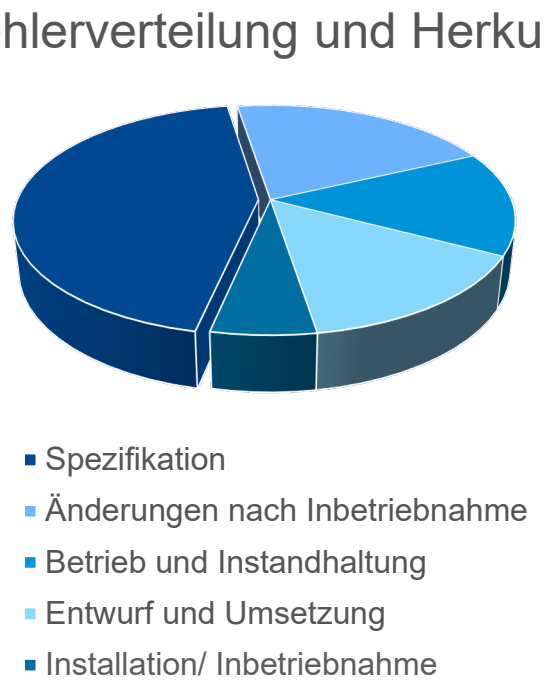
   Es liegen keine Informationen darüber vor, wie das System funktioniert.



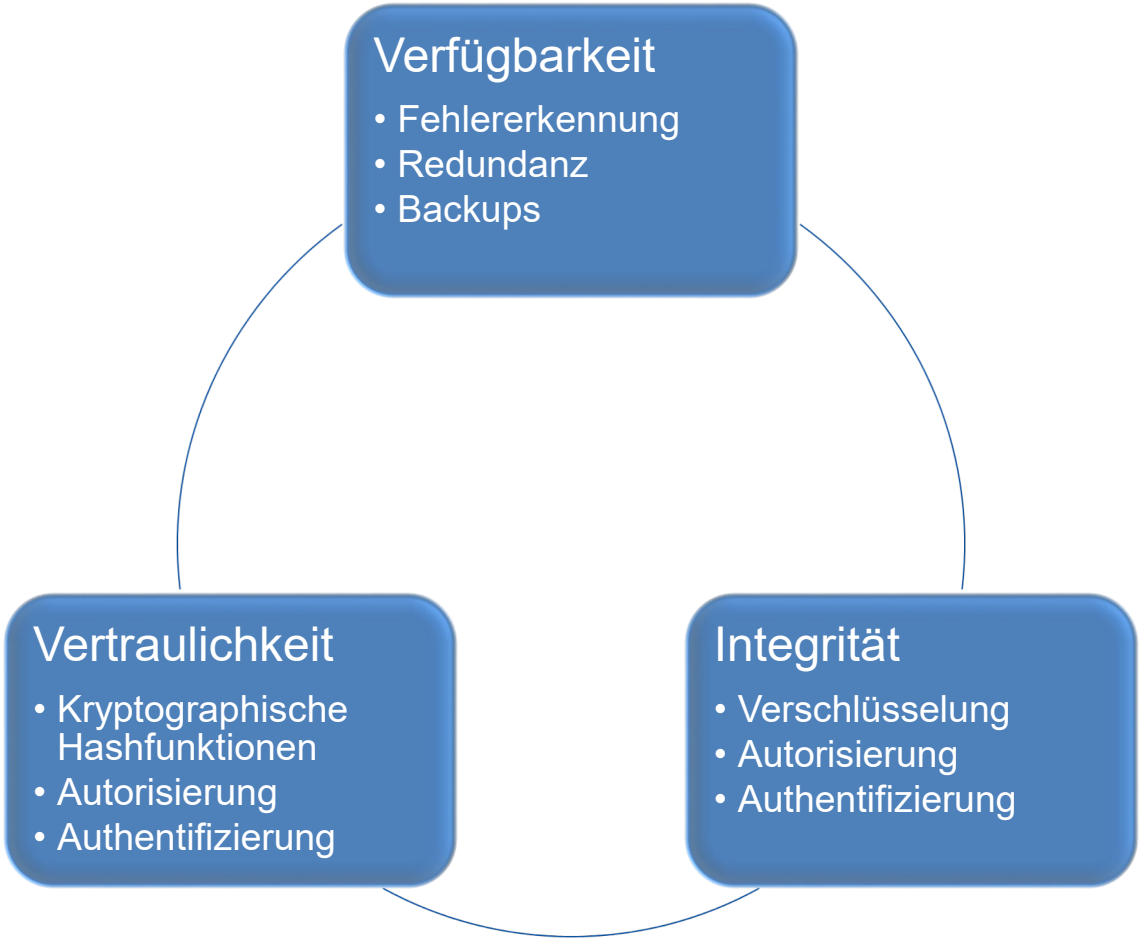
5. Zuverlässigkeit und Robustheit des Modells

Spezifikation

- Formulierung des Problems
  - Anforderung an einen Roboter: keinen Menschen verletzen
    - Anweisung: Greife NICHT nach einem menschlichen Wesen
    - Anweisung: Greife NUR nach einem definierten Objekt
- Übertragung der Spezifikation in das System
  - Klassisches System:
    - Spezifikation wird vom Entwicklungsteam interpretiert
  - ML-System:
    - Das mentale Konzept der Spezifikation muss implizit durch die Daten beschrieben werden
    - Der Trainingsprozess findet nicht immer die bestmögliche Lösung oder wird durch Datenverzerrungen beeinflusst



6. Security


Verfügbarkeit
• Fehlererkennung
• Redundanz
• Backups
Vertraulichkeit
• Kryptographische Hashfunktionen
• Autorisierung
• Authentifizierung
Integrität
• Verschlüsselung
• Autorisierung
• Authentifizierung

7. System-Hardware

- Es müssen zwei Systeme betrachtet werden:
  - Trainingssystem:
    - Training erfordert viel Rechenleistung
    - Cloud-Systeme, Edge-Systeme, GPU-Cluster
  - Applikationssystem
    - Anwendung des fertigen Modells erfordert meist weitaus weniger Rechenleistung
    - Edge-Systeme, GPUs, Embedded-Systeme
  - Asymmetrie zwischen Trainingsphase und Applikationsphase
    - Verschiedene Speicherverwaltung, Speicherarchitektur sowie Speichergröße
    - Verschiedene Programmiersprachen
  - ➢ Übersetzungsfehler

8. Technologische Ausgereiftheit

- Bei neuen Technologien sind meist noch keine ausreichenden Informationen über das tatsächlich vorhandene Risiko verfügbar
- Bei alten Technologien sinkt oft das Risikobewusstsein mit der Zeit

1. **Aufkommend:** Wird für einen möglichen zukünftigen Einsatz erforscht und erprobt.
2. **Strategisch:** Ist voraussichtlich erst mittel- bis langfristig einsatzfähig.
3. **Begrenzt:** Ist für die Umsetzung einer begrenzten Anzahl an Anwendungen bereits einsatzfähig.
4. **Bevorzugt:** Wird zur Umsetzung der meisten Anwendungen bereits bevorzugt.
5. **Aktuell:** Wird derzeit unterstützt und verwendet.
6. **Außer Dienst:** Kurz davor nicht mehr verwendet zu werden.

IFA
Institut für Arbeitsschutz der
Deutschen Gesetzlichen Unfallversicherung

Vielen Dank
für Ihre Aufmerksamkeit.

Dr. André Steimers
DGUV - IFA - Abteilung 5 „Unfallprävention: Digitalisierung – Technologien“

Telefon: +49 30 13001 3539
eMail: andre.steimers@dguv.de