DGUV
Deutsche Gesetzliche Unfallversicherung
Spitzenverband

Digitalisierung & KI
Standortbestimmung Künstliche Intelligenz

DGUV Fachgespräch “Künstliche Intelligenz”
8./9. November 2022

Dietmar Reinert, Moritz Schneider & André Steimers

07.12.2022

Überblick

Worüber reden wir, wenn wir KI meinen?

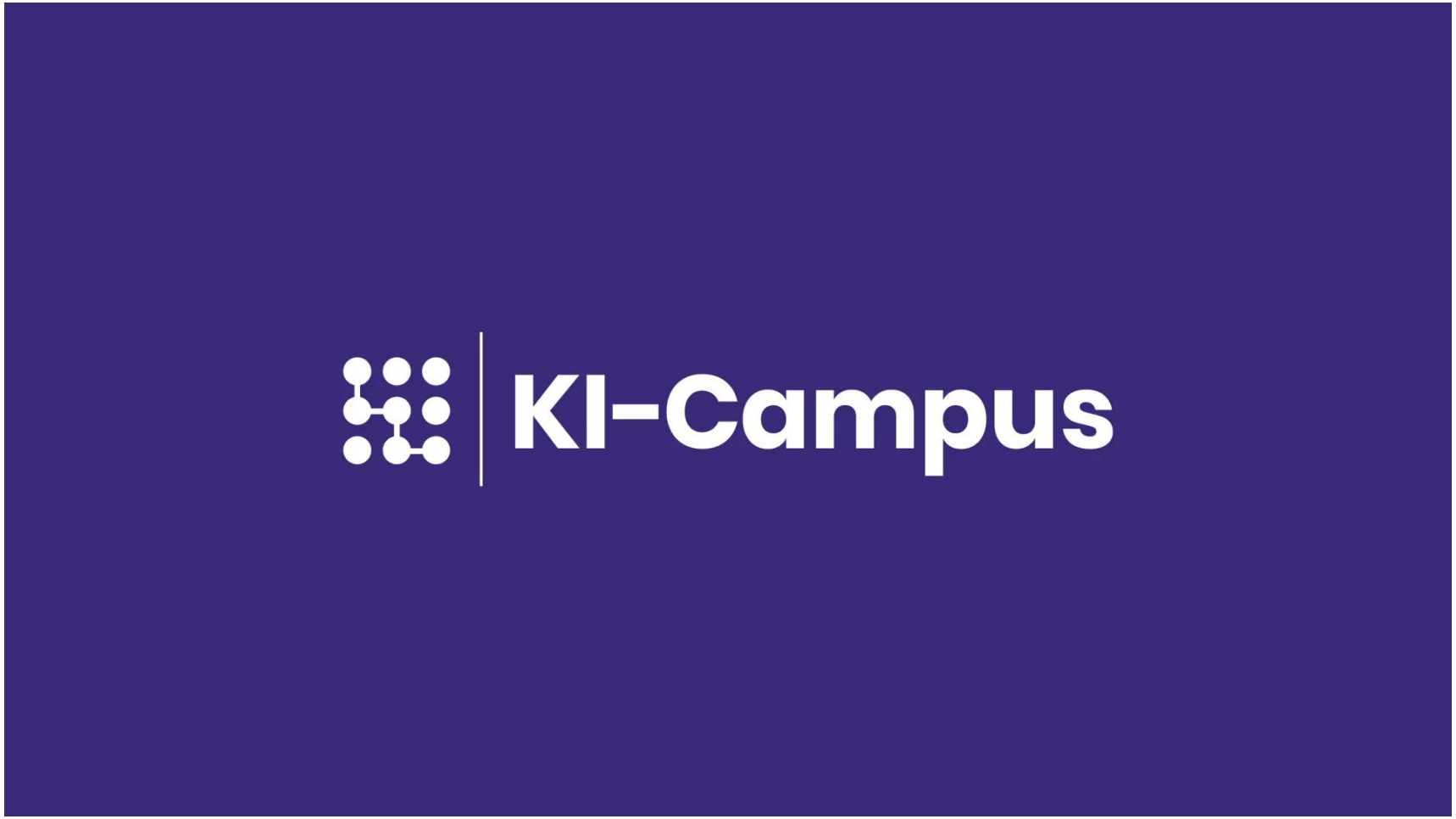
Warum ist KI für die Unfallversicherung wichtig?

KI reduziert sich nicht nur auf Neuronale Netze.

Bedeutung der vertrauenswürdigen KI

Anwendungen im IFA

Was geht derzeit nicht



https://ki-campus.org

Alexander Waldmann beschreibt hier die starke KI

Diese ist aber noch Utopie.

Derzeit können die Systeme nur schwache KI

Aber es gibt doch…

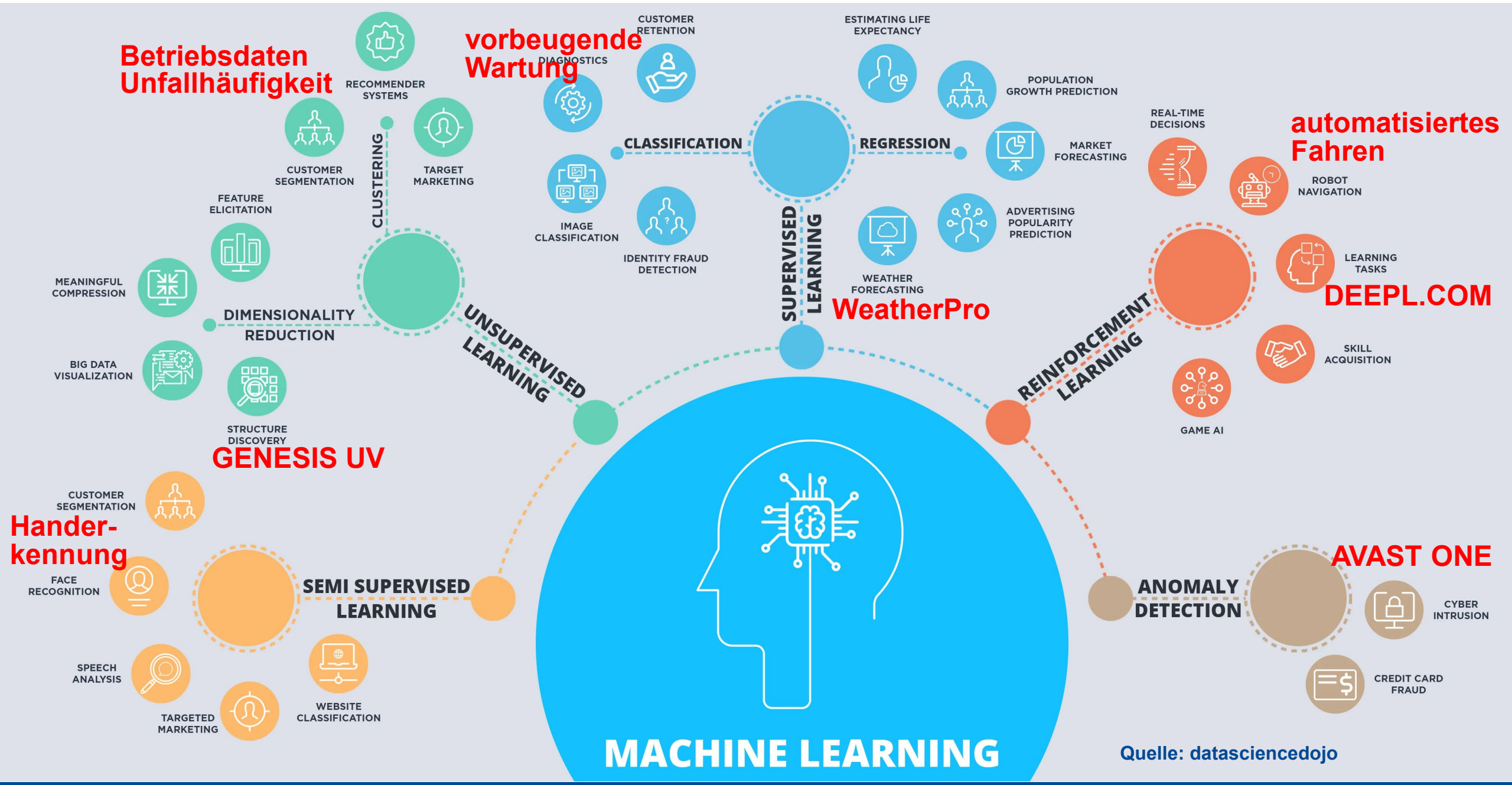
Bildquelle: dpa

https://www.magentamusik.de/beethoven-10-sinfonie

Tim Höttges, Vorstandsvorsitzender der Deutschen Telekom AG, ab 2:30

3. und 4. Satz ab 17:12

Das ist keine starke KI

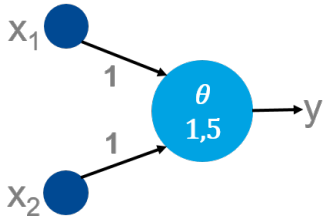


Quelle: datasciencedojo

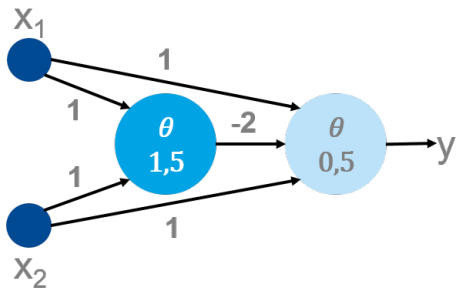
Welche Werkzeuge stellt die KI zur Verfügung?

Quelle: TheInsaneApp

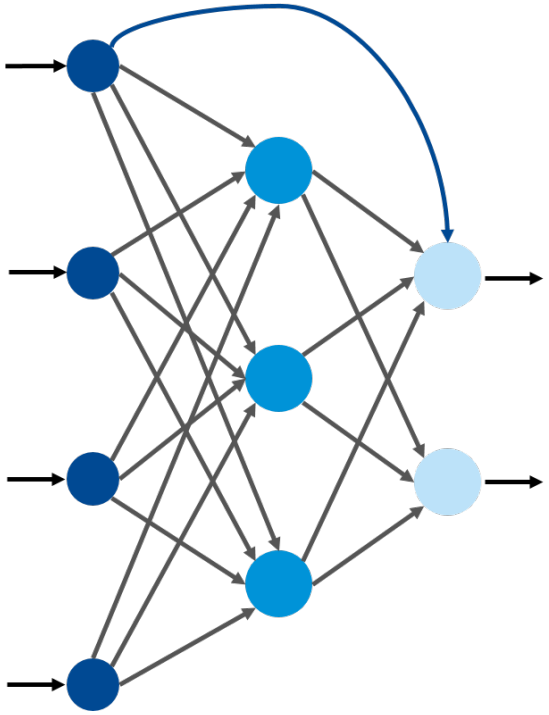
Was macht ein neuronales Netzwerk?



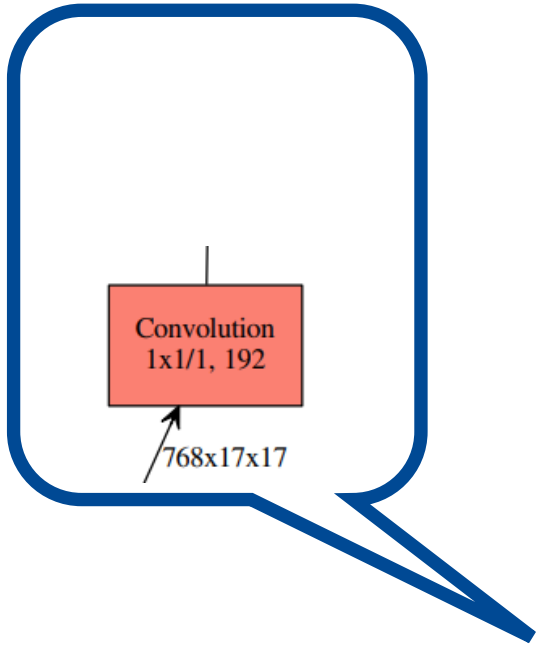
1 Neuron



2 Neurone



5 Neurone



768x17x17 Neurone

Nachteile der tiefen neuronalen Netze

Eine Kuh in anderer Umgebung wird zum Hund!

Quelle: https://www.youtube.com/watch?v=HuorfODPjqA

- Neuronale Netze sind wegen ihrer Komplexität eine Black-Box
- Man findet manchmal ein sehr seltsames Verhalten
- Das muss erklärt werden, wenn die Systeme im Bereich von Sicherheit und Gesundheit eingesetzt werden sollen.

Die goldene Regel

Problemquellen, die allein mit gesundem Menschenverstand zu finden und zu erklären sind:

- Probleme mit den **Daten**
- Probleme mit dem **Algorithmus**
- Probleme im **Training**: Unter- und Überanpassung

Bei der Suche nach Problemen immer **Vergleiche aus dem echten Leben** heranziehen, um die abstrakten Zahlen mit ihren versteckten Problemen besser zu erfassen.

- ML-Anwendung anhand ihrer **Leistung in neuen untrainierten Situationen** beurteilen
- Rate von **Klassifikationsfehlern** anhand der Einordnung in die **Konfusionsmatrix** bestimmen. Sind die “falsch positiven” und “falsch negativen” Ergebnisse zu hoch, muss die ML-Lösung angepasst werden.
- **Visualisierung von Daten** nutzen, um einen besseren Überblick über die ML-Anwendung zu erhalten und eventuelle Fehler zu erkennen.

Ethische und sicherheitstechnische Aspekte

Vertrauenswürdigkeit der KI

Abhängig von den Risikoquellen des gewählten KI-Verfahrens

1. Fairness
2. Privacy
3. Automatisierungsgrad und Kontrolle

Ethische Aspekte

4. Komplexität der Aufgabe und Verwendungsumgebung
5. Grad der Transparenz und Erklärbarkeit
6. Security
7. System-Hardware
8. Technologische Ausgereiftheit

Zuverlässigkeit und Robustheit

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

Maßnahmen

1. Analyse der Anwendung auf Möglichkeit der Diskriminierung
2. Analyse der genauen Diskriminierungsszenarien
3. Analyse zur Identifikation seltener Fälle
4. Sicherstellung der Datenqualität
   - Vollständigkeit der Daten
   - Diversität der Daten
   - Ausgewogenheit der Klassen innerhalb der Daten
   - Anreicherung der Trainingsdaten
5. Anwendung von Fairness Metriken zur Evaluation (z. B. statistical parity, predictive rate parity, equalized odds)

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

System	Grad der Automatisierung	Grad der Kontrolle	Kommentare
Autonomie	Autonom	Human out of the loop	Das System ist in der Lage, seinen Betriebsbereich oder seine Ziele ohne Eingreifen, Kontrolle oder Aufsicht von außen zu ändern.
Heteronomie	Vollautomatisiert	Human on the loop Human out of the loop	Das System ist in der Lage, seine gesamte Aufgabe ohne Eingriffe von außen zu erfüllen.
	Hochautomatisiert	Human on the loop	Das System führt Teile seiner Aufgaben ohne externen Eingriff aus.
	Bedingte Automatisierung	Human on the loop	Nachhaltige und spezifische Leistung eines Systems, wobei ein externer Agent bereit ist, bei Bedarf einzuspringen.
	Teilautomatisierung	Human in the loop	Einige Teilfunktionen des Systems sind vollständig automatisiert, während das System unter der Kontrolle eines externen Agenten bleibt.
	Assistenzfunktion	Human in the loop	Das System unterstützt den Bediener.
	Keine Automatisierung	Human in the loop	Der Bediener hat die volle Kontrolle über das System.

4. Komplexität der Aufgabe und Verwendungsumgebung

- **Grundsätzlich hilft die Goldene Regel: Das Lernen müssen wir Menschen kritisch begleiten. Derzeit ist es für sicherheitsrelevante Systeme nicht sinnvoll, das Lernen alleine der Maschine zu überlassen!**
- Aufgabenstellung und Umgebung sind in diesem Zusammenhang zu betrachten
- Die Datenqualität beim Lernprozess ist sehr wichtig
- Folgende Kriterien sind zu beachten: Genauigkeit, Präzision, Vollständigkeit, Repräsentativität, Konsistenz, Relevanz, Skalierbarkeit, Kontextabdeckung, Übertragbarkeit, Latenzzeit, Aktualität, Identifizierbarkeit, Überprüfbarkeit, Glaubwürdigkeit

5. Grad der Transparenz und Erklärbarkeit

Quelle: Andre Steimers

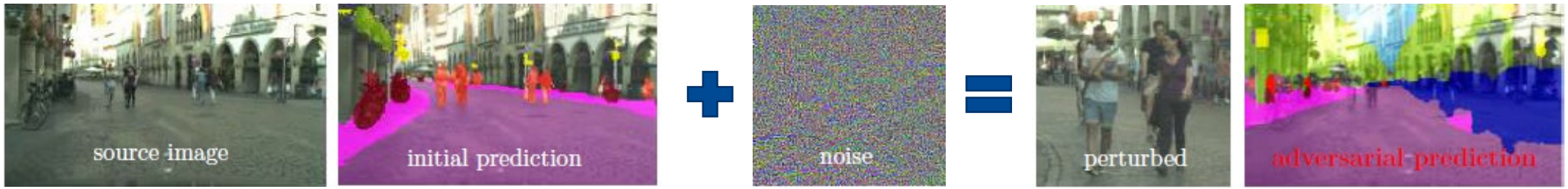
Hier geht es darum, ob man die Entscheidungen der KI nachvollziehen kann. Es gibt sogenannte White Box verfahren wie: Entscheidungsbäume, Entscheidungsregeln, Random Forrests, (Fuzzy) Regellernen oder Induktive Logische Programmierung. Diese können aber auch sehr komplex werden und müssen dann wie jede komplexe Software behandelt werden. Zu den Black-Box Verfahren gehören die Neuronalen Netz und die sind auch für Entwickler intransparent, weil die Entscheidungsfunktion auf komplexer, hochgradig nicht-linearer Verrechnung basiert. Bei diesen Verfahren benötigt man Ansätze wie Layer-Wise Relevance Propagation (LRP) oder Local Interpretable Model-agnostic Explanations (LIME) bei denen man die Gründe für die Entscheidung sichtbar und damit nachvollziehbar macht. Dazu ist das Tutorial Erklärbares Maschinelles Lernen für Ingenieurwissenschaften im KI-Campus sehr hilfreich. Beim LRP geht man grob gesagt vom Ergebnis des Neuronalen Netzes aus und rechnet zurück auf die Eingangsdaten, die zu dem Ergebnis geführt haben.
Die Farben der Umrandungen im Bild machen deutlich, wo ein Boot nicht mehr als Boot und eine Hund nicht mehr als Hund erkannt wird. Das gilt es dann über LRP z.B. nachzuvollziehen.
Diese Arbeit kann Teil einer Prüfung eines KI-Systems sein.

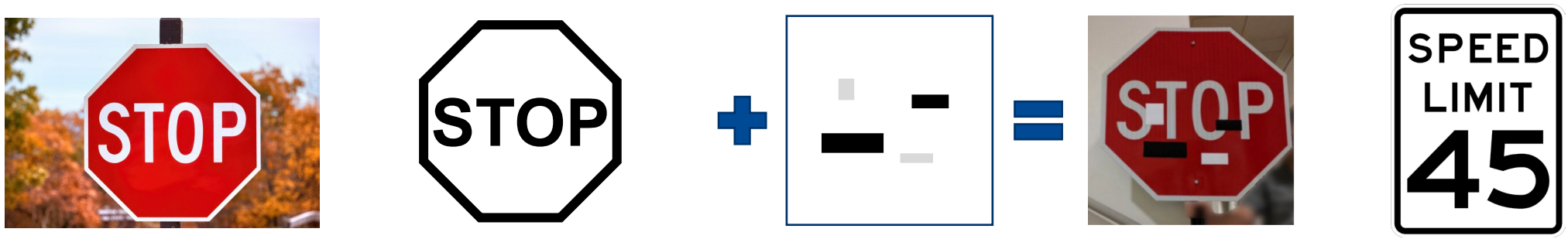
6. Security

Adverserial Attacks

Die Daten von KI liegen oft irgendwo in einer Cloud. Damit werden Angriffe von Außen möglich, sodass Daten über gezielte Fehlinformation zu falschen Ergebnissen führen können (siehe Beispiele aber auch Folie 10 mit dem Schwein, dass als Flugzeug und der Kuh, die als Hund klassifiziert wurde). Die Security wird also bei der Anwendung von KI, auch wenn es gar nicht um funktionale Sicherheit sondern um Datenauswertungen geht, sehr wichtig. Das ist auch der Grund, warum wir im IFA das Thema KI auf die Tagesordnung im Zusammenhang mit Industrie 4.0 gesetzt haben. Siehe auch Fachbereich AKTUELL FBHM-102: Sachgebiet Maschinen, Robotik und Fertigungsautomation Safety und Security in der vernetzten Produktion. Auch der Prüfgrundsatz GS-IFA-M24: „Grundsätze für die Prüfung und Zertifizierung von Security-Aspekten in der funktionalen Sicherheit von industriellen Automatisierungssystemen“ kann da wichtig sein. Das IFA hat unter seinen Fachinfos eine eigene Seite zu „Industrial Security“. Die Security wird durch die KI wichtiger aber nicht grundsätzlich anders.

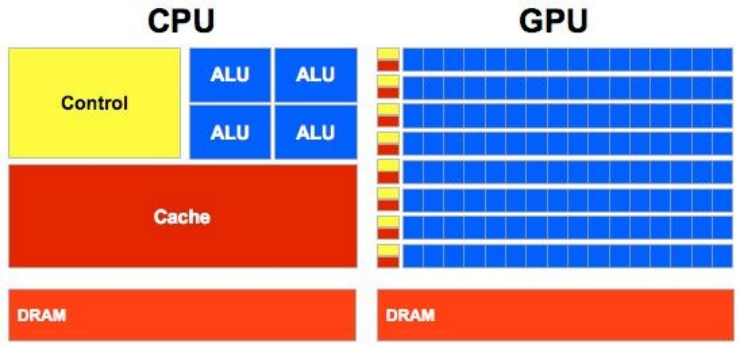
- Einem gültigen Modell werden gestörte Eingangsdaten geliefert um dieses zu täuschen





Quellen: Koopman et. al., Challenges in autonomous vehicle testing and validation, SCAV 17, 2017
Eykholt et. al., Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR, 2018

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

Grundsätzlich gilt für die Hardware, was wir aus der funktionalen Sicherheit kennen. Das Trainingssystem ist sozusagen ähnlich wie der Compiler in Standardanwendungen. Allerdings wird zum Training oft viel leistungsfähigere Hardware verwendet (GPU-Cluster) als zu konkreten Anwendung, z.B. in einem eingebetteten System. Eigentlich wäre man auf der sicheren Seite, wenn man das Trainingssystem wie einen Compiler validiert hätte. Das dürfte aber schwierig sein, weil die Anwendung solcher Hardware in der Regel nicht in Sicherheitsanwendungen unterwegs sind und dort auch kaum Erfahrungen mitbringen. Was die Anwendungshardware betrifft, sollten die Maßnahmen aus der funktionalen Sicherheit angewendet werden.

8. Technologische Ausgereiftheit

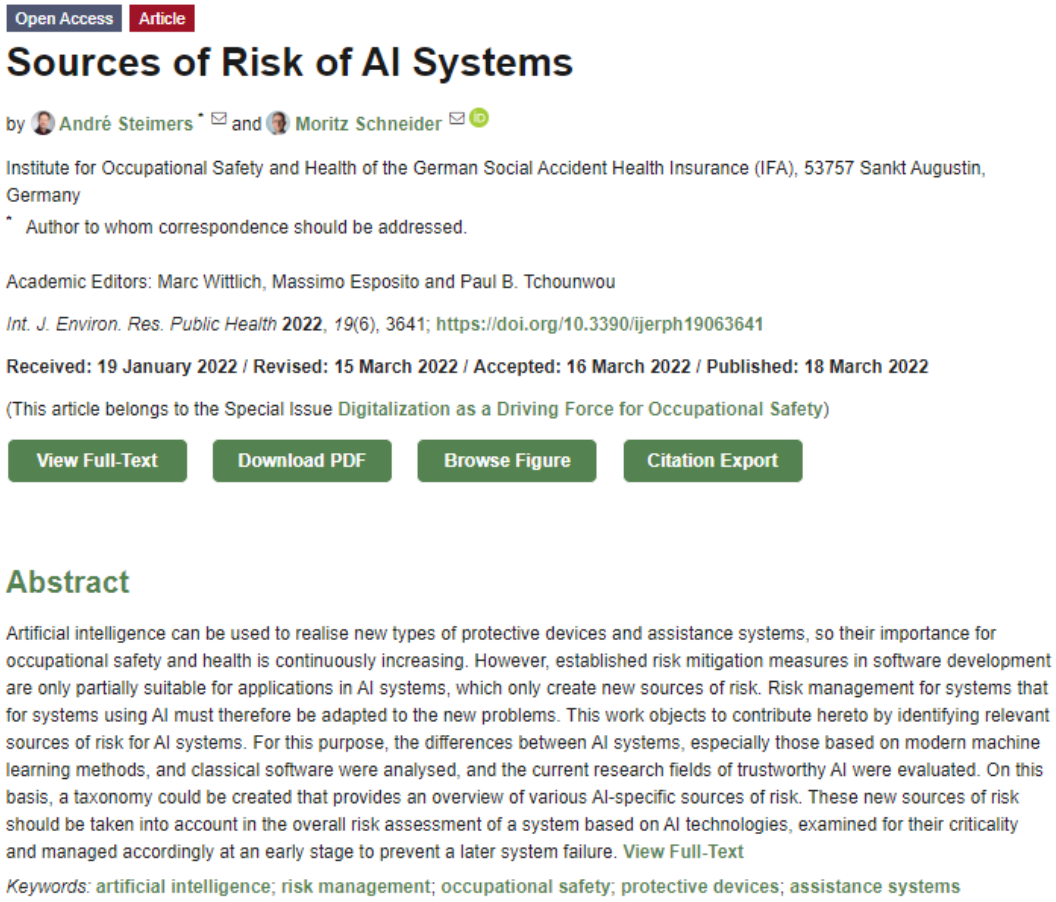
- Bei neuen Technologien sind meist noch keine ausreichenden Informationen über das tatsächlich vorhandene Risiko verfügbar
- Bei alten Technologien sinkt oft das Risikobewusstsein mit der Zeit

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.

Weitere Informationen

- **Veröffentlichung:**

- Steimers A. and Schneider M.
- **Sources of Risk of AI Systems**
- International Journal of Environmental Research and Public Health. 2022; 19(6):3641.
- https://doi.org/10.3390/ijerph19063641



Open Access Article

Sources of Risk of AI Systems

by André Steimers * and Moritz Schneider

Institute for Occupational Safety and Health of the German Social Accident Health Insurance (IFA), 53757 Sankt Augustin, Germany

* Author to whom correspondence should be addressed.

Academic Editors: Marc Wittlich, Massimo Esposito and Paul B. Tchounwou

Int. J. Environ. Res. Public Health **2022**, *19*(6), 3641; https://doi.org/10.3390/ijerph19063641

Received: 19 January 2022 / Revised: 15 March 2022 / Accepted: 16 March 2022 / Published: 18 March 2022

(This article belongs to the Special Issue Digitalization as a Driving Force for Occupational Safety)

View Full-Text | Download PDF | Browse Figure | Citation Export

Abstract

Artificial intelligence can be used to realise new types of protective devices and assistance systems, so their importance for occupational safety and health is continuously increasing. However, established risk mitigation measures in software development are only partially suitable for applications in AI systems, which only create new sources of risk. Risk management for systems that for systems using AI must therefore be adapted to the new problems. This work objects to contribute hereto by identifying relevant sources of risk for AI systems. For this purpose, the differences between AI systems, especially those based on modern machine learning methods, and classical software were analysed, and the current research fields of trustworthy AI were evaluated. On this basis, a taxonomy could be created that provides an overview of various AI-specific sources of risk. These new sources of risk should be taken into account in the overall risk assessment of a system based on AI technologies, examined for their criticality and managed accordingly at an early stage to prevent a later system failure. View Full-Text

Keywords: artificial intelligence; risk management; occupational safety; protective devices; assistance systems

Beispiel: eigene Entwicklung - SRS

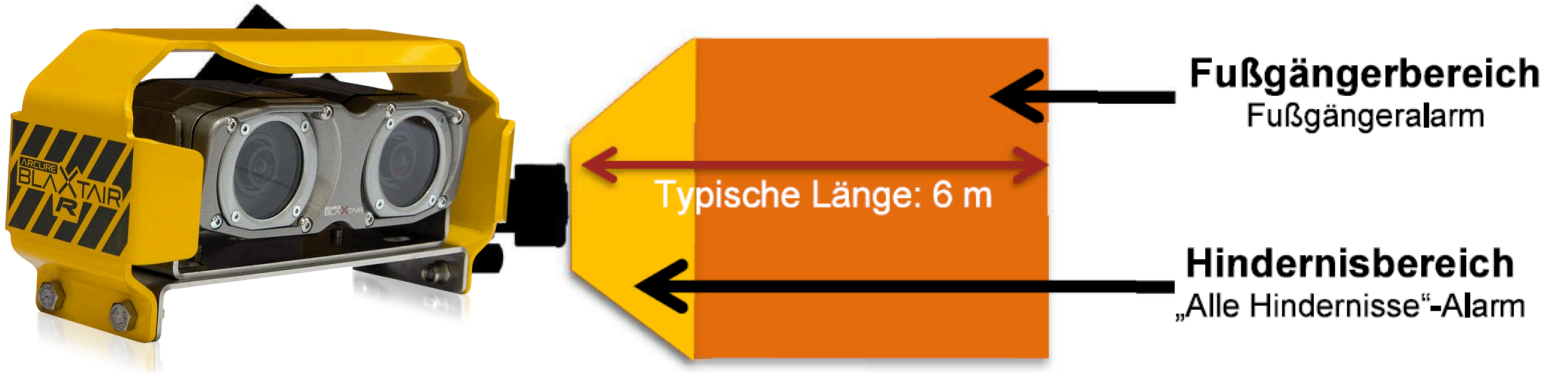
Einsatz von maschinellem Lernen für (Beinahe-) Sturzdetektion zur Prävention von Stolper-, Rutsch- und Sturzunfällen

- **DGUV-gefördertes Forschungsprojekt (FP 470-SRS)**
  - Entwicklung zusätzlicher Trainingselemente zur Prävention von Stolper-, Rutsch- und Sturzunfällen unterstützt durch den Einsatz von virtueller Realität am Beispiel von Unternehmen der Stahlerzeugung und der Post- und Paketzustellung (ENTRAPon)

- **Unsere Ziele**
  - Gewinnung standardisierter kinematischer Daten von „Beinahestürzen“
  - Entwicklung Algorithmen zur Erkennung von Beinahestürzen
  - Langfristig: Einfaches Messsystem zur Erkennung und Quantifizierung von Beinahestürzen in der Berufspraxis

Beispiel: Assistenzsystem Blaxtair

- **Auftraggeber:** BGHW
- **Kooperationspartner:** Jungheinrich, Arcure Blaxtair
- **Projekt:** Entwicklung einer intelligenten Kamera, die in Echtzeit eine Person von einem anderen Hindernis unterscheiden und den Fahrer im Gefahrenfall warnen kann.



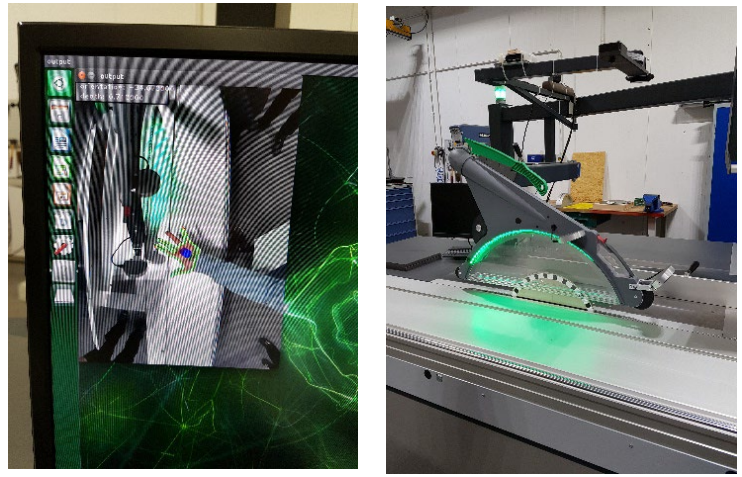
Quelle: https://blaxtair.com/de

Beispiel: Assistenzsystem für Formatkreissägen

Auftraggeber: BGHM

Kooperationspartner: Altendorf

Projekt: Assistenzsystem das die Bewegungen der Hände eines Operators erkennt und die Maschine, in Abhängigkeit von deren Position, Richtung und Geschwindigkeit, in einen sichereren Zustand versetzt. HAND GUARD erkennt Gefahrenpotentiale basierend auf einer speziellen KI-Handerkennungssoftware lange vor Eintritt einer möglichen Verletzung. Bei einer akuten Gefährdung von Fingern oder der Hand senkt sich das Sägeblatt innerhalb einer Viertelsekunde durch Schnellabsenkung des Sägeaggregats ab, welches unter den Bearbeitungstisch abtaucht.



Quelle: Altendorf Group

DGUV
Deutsche Gesetzliche Unfallversicherung
Spitzenverband

Vielen Dank
für Ihre Aufmerksamkeit.

Prof. Dr. Dietmar Reinert; Moritz Schneider
DGUV - IFA

Telefon: +49 30 13001 3000; 3164
eMail: dietmar.reinert@dguv.de;
moritz.schneider@dguv.de

Prof. Dr. Andre Steimers
RheinAhrCampus der Hochschule Koblenz

Telefon: +49 2642 932 215
eMail: steimers@hs-koblenz.de