

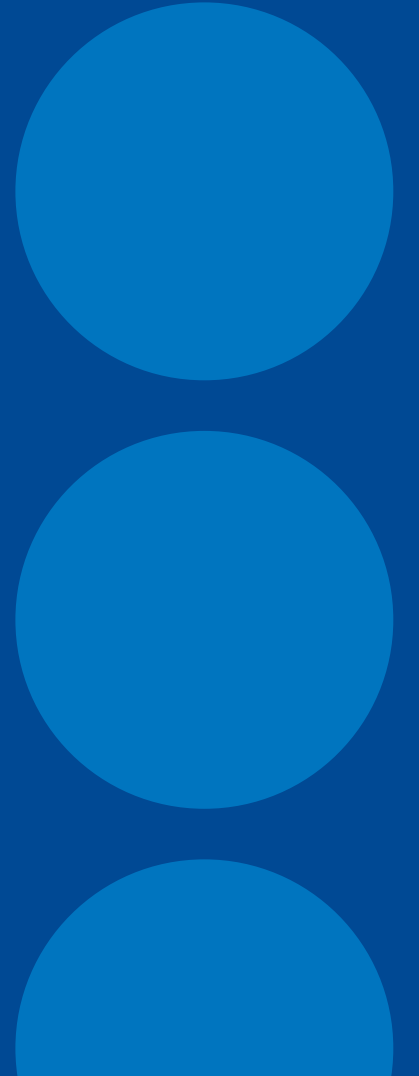
Characteristics of trustworthy artificial intelligence

5th PEROSH Research Conference &
Anniversary event

7th of September 2023

Dietmar Reinert, IFA, Germany

27.09.2023



Overview

Will AI take over the world leadership?

Why should we use AI for OSH?

Machine learning is more than neural networks.

Trustworthy AI as a chance

Examples using AI for OSH in IFA

Is AI a threat?



AI has elements which could be dangerous:

Prejudices could influence decisions.

Wrong interpretation of data could delay reactions.

Misuse could attack social media.

But there is no strong AI available today.

But what's about... **ChatGPT:**

Question: Are you a trustworthy artificial intelligence?

As an AI language model, I am designed to provide accurate and informative responses based on the knowledge I have been trained on. However, as with any AI system, my trustworthiness ultimately depends on the accuracy and completeness of the information I have been trained on, as well as the limitations of my programming.

It's important to note that I do not have personal beliefs or motivations, nor do I have the ability to make decisions or take actions outside of my programming. Therefore, it's important to approach my responses with a critical mindset and to cross-reference my answers with other sources to ensure accuracy.

Overall, while I strive to be a trustworthy AI, it's important to remember that all AI systems have limitations and potential biases and should be used in conjunction with human judgement and oversight.

This is not the answer of a strong AI!!!

But what's about...

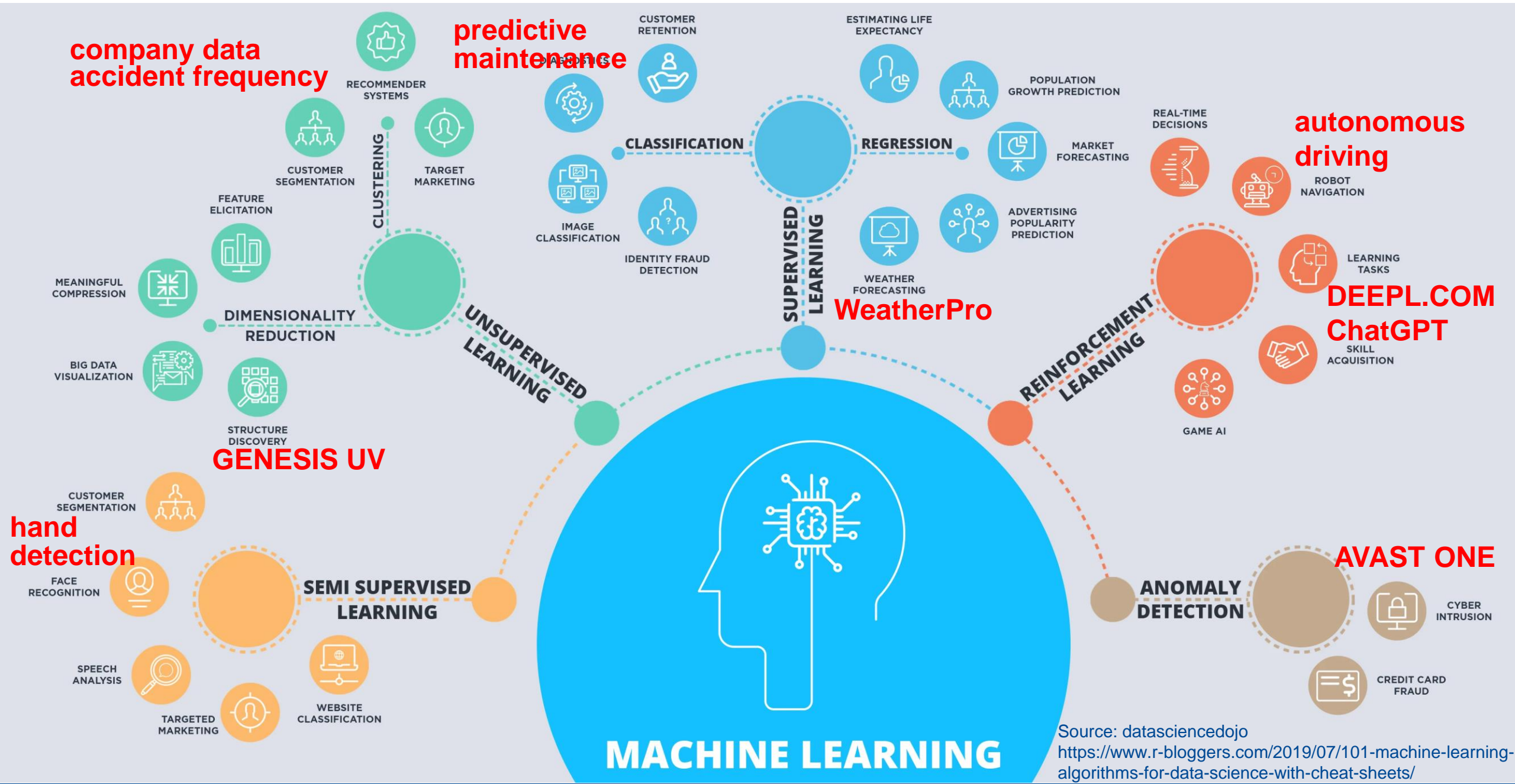


<https://www.magenta-musik.de/beethoven-10-sinfonie>

Tim Höttges, CEO of Deutsche Telekom, starting at 2:30

3. and 4. Satz starting at 17:12

This is a summary of the old Beethoven, not really new.



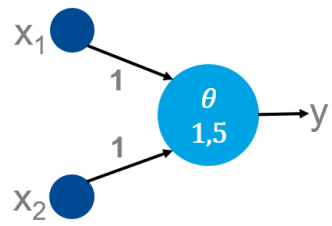
Source: datasciencedojo
<https://www.r-bloggers.com/2019/07/101-machine-learning-algorithms-for-data-science-with-cheat-sheets/>

Which algorithms we have for machine learning?

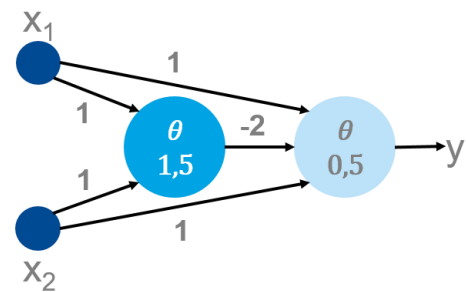


Source: TheInsaneApp
<https://twitter.com/crigatuso/status/1376094307413712898>

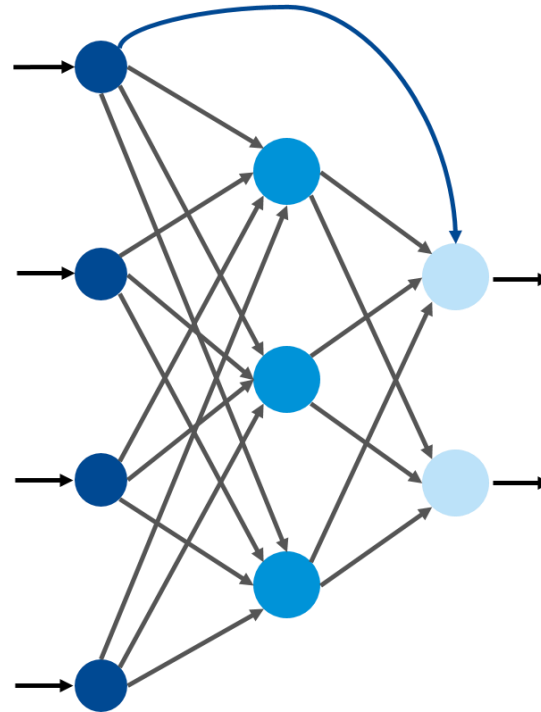
How does a neural network work?



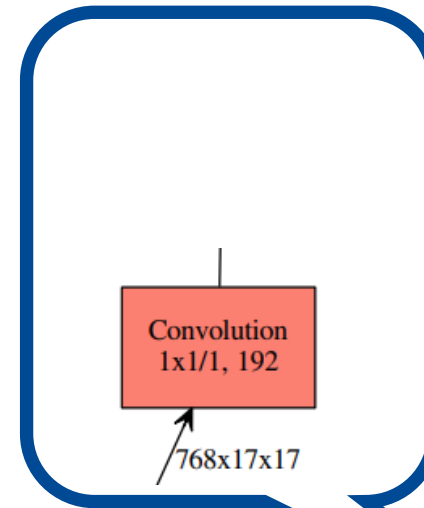
1 Neuron



2 Neurone

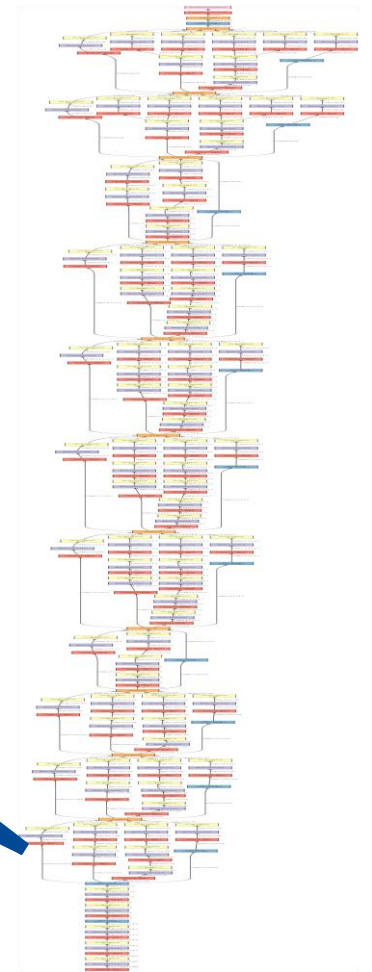


5 Neurone



768x17x17 Neurone

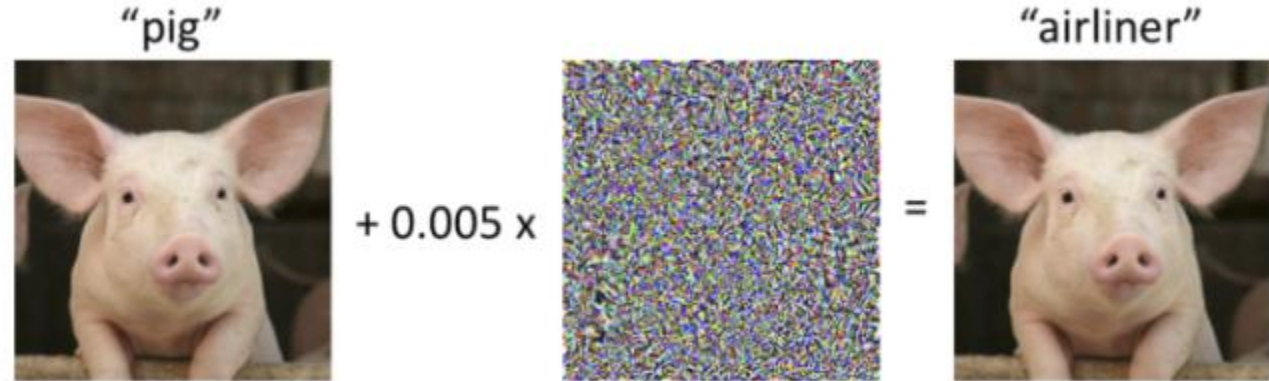
Inception Net V3



Source: www.edge-ai-vision.com/

Deficits of the deep neural networks

A pig with a little noise is recognized as an airliner!



Source: https://gradientscience.org/intro_adversarial/



A cow in unusual environment is labelled as a dog!

Source: <https://www.youtube.com/watch?v=HuorfODPjqA>

The golden rule!

Sources of problems that can be found and explained with common sense alone:

- Problems with the **data**
- Problems with the **algorithm**
- Problems during the **training**: Under- und Over-Fitting

When looking for problems, always use real-life comparisons to better grasp the abstract numbers with their hidden problems.

- Evaluate ML application based on its performance in new untrained situations.
- Determine rate of classification errors based on classification in the confusion matrix. If the "false positives" and "false negatives" are too high, the ML solution needs to be adjusted.
- Use visualization of data to get a better overview of the ML application and identify possible errors.

Ethical and safety-related aspects

Trustworthiness of AI

Depending on the sources of risk of the selected AI procedure

1. Fairness
2. Privacy
3. Degree of automation and control



Ethical aspects

4. Complexity of the task and usage environment
5. Degree of transparency and explainability
6. Security
7. System hardware
8. Technological maturity



Reliability and robustness

1. Fairness



Source: www.jobrapido.com



Quelle: www.embedica.ai/fairness-2



Source: Joy Buolamwini, M.I.T. Media Lab

- **Recruiting-Tool...**
...discriminates against women
- **Historical Bias**
ML model can learn negative correlation as men were often systematically favoured in the past
- **Face recognition**
poorer performance among people of colour
- **Data Bias**
Underrepresented groups in the training data lead to higher error rates of these groups in the ML model

2. Privacy

EU General Data Protection Regulation 2016/679

Article 5 Principles...	Personal data shall be:
lawfulness, fairness and transparency	„processed lawfully, fairly and in a transparent manner in relation to the data subject.“
purpose limitation	„collected for specified, explicit and legitimate purposes...“
data minimisation	„adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.“
accuracy	„accurate and, where necessary, kept up to date...“
storage limitation	„kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; ...“
integrity and confidentiality	„... processed in a manner that ensures appropriate security of the personal data, ...“

3. Degree of automation and control

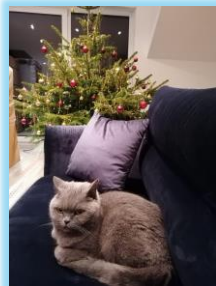
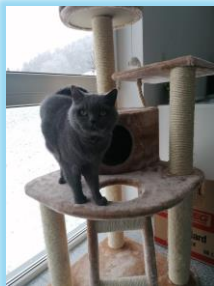
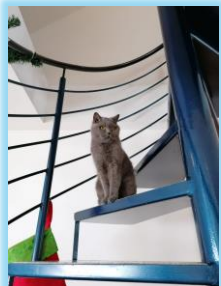
System	Degree of automation	Degree of control	comments
Autonomy	Autonom	Human out of the loop	The system is capable of changing its operating range or objectives without outside intervention, control or supervision.
Heteronomy	Fully automated	Human on the loop Human out of the loop	The system is able to perform its entire task without external intervention.
	Highly automated	Human on the loop	The system performs parts of its tasks without external intervention.
	conditional automation	Human on the loop	Sustained and specific performance of a system, with an external agent ready to step in when needed.
	Semi-automated	Human in the loop	Some sub-functions of the system are fully automated, while the system remains under the control of an external agent.
	Assistance	Human in the loop	The system supports the operator.
	not automated	Human in the loop	The operator has full control over the system.



4. Complexity of the task and usage environment



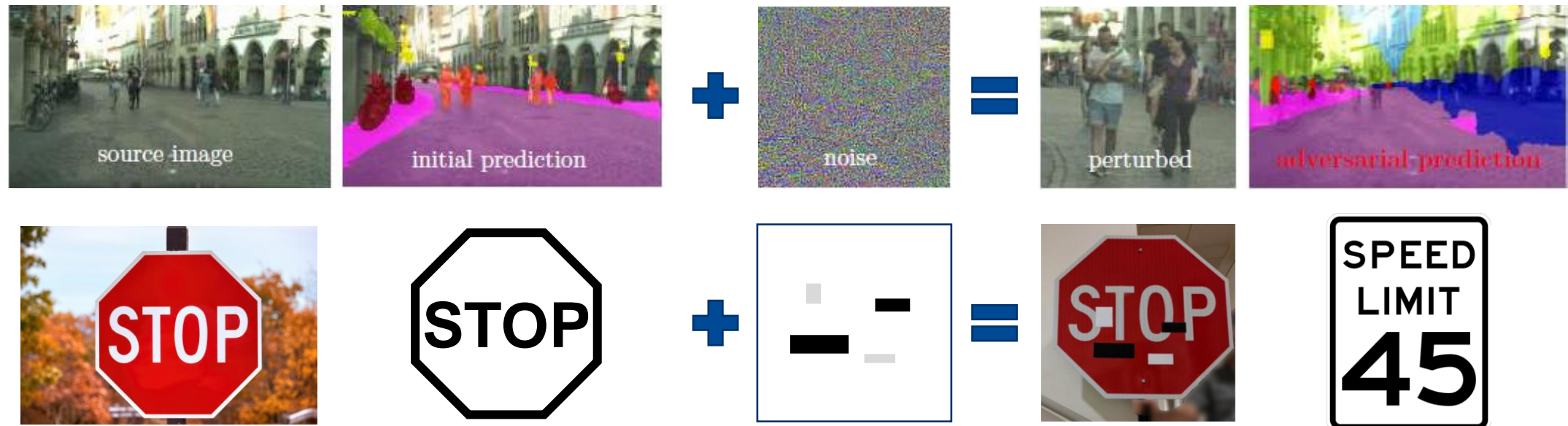
5. Degree of transparency and explainability



6. Security

Adversarial Attacks

- A valid model is supplied with disturbed input data in order to deceive it.

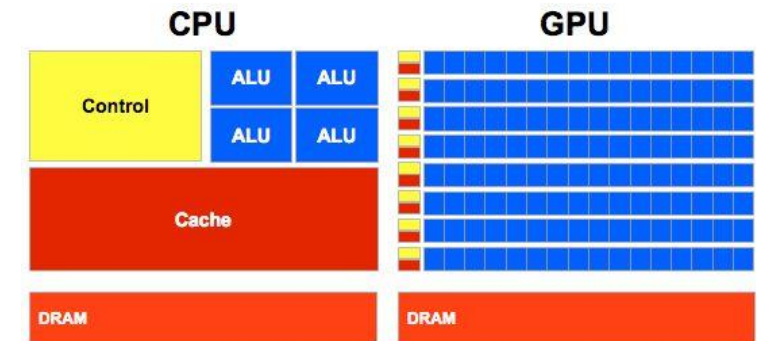


Sources: Koopman et. al., Challenges in autonomous vehicle testing and validation, SCAV 17, 2017
 Eykholt et. al., Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR, 2018

7. System-Hardware

Two systems need to be considered:

- Training system:
 - Training requires a lot of computing power
 - Cloud systems, edge systems, GPU clusters
- Application system
 - Application of the finished model usually requires much less computing power
 - Edge systems, GPUs, embedded systems
- Asymmetry between training phase and application phase
 - Different memory management, memory architecture and memory size
 - Different programming languages



8. Technological maturity






- In the case of new technologies, there is usually not yet sufficient information available about the actual risk involved.
 - With old technologies, risk awareness often decreases over time
1. **Emerging:** Being researched and tested for possible future use.
 2. **Strategic:** Likely to be operational only in the medium-to-long term.
 3. **Limited:** Already operational for the implementation of a limited number of applications.
 4. **Preferred:** Already preferred for the implementation of most application
 5. **Current:** Currently supported and in use.
 6. **Out of service:** Is on the verge of no longer being used..

further information

- recent publication:
- 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

Example: Research project – Trips, slips and falls

Use of machine learning for (near) fall detection for prevention of trip, slip and fall accidents

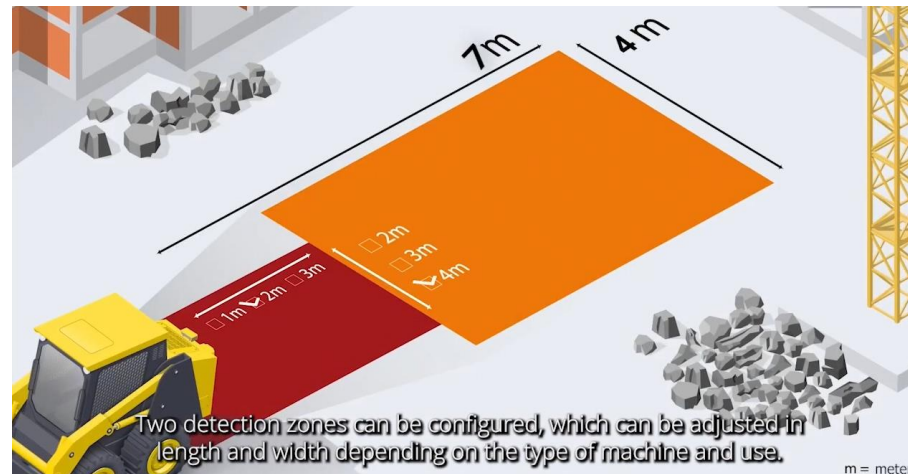
- **DGUV-funded research with partners (FP 470-SRS)**
 - Development of additional training elements for the prevention of tripping, slipping and falling accidents supported by the use of virtual reality, using the example of companies in steel production and postal and parcel delivery (ENTRAPon).
- **Our goals**
 - Obtaining standardised kinematic data from "near fall detection".
 - Development of algorithms for the detection of "near fall detection".
 - Long-term: Simple measurement system to detect and quantify near falls in professional practice.



Source: DGUV

Example: Assistance system Blaxtair

- **Initiator:** German Social Accident Insurance Institution for the trade and logistics industry
- **Cooperation partner:** Jungheinrich, Arcure Blaxtair
- **Project:** Development of an intelligent camera that can distinguish a person from another obstacle in real time and warn the driver in case of danger.



Source: <https://blaxtair.com/en>



Example: Assistance system for sliding table saws

Initiator: German Social Accident Insurance Institution for the woodworking and metalworking industries

Cooperation partner: Altendorf group

Projekt: Assistance system that recognises the **movements of an operator's hands** and sets the machine to a safer state depending on their **position, direction and speed**. HAND GUARD detects potential hazards based on **special AI hand detection software** long before a possible injury occurs. In the event of an **acute risk to fingers or the hand**, the **saw blade lowers within a quarter of a second** by rapidly lowering the saw unit, which dives under the machining table.



Source: Altendorf group

<https://www.altendorf-handguard.com/en/>

Thank you for your attention.

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

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

Prof. Dr. Andre Steimers
RheinAhrCampus of Hochschule Koblenz
University of Applied Sciences

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

