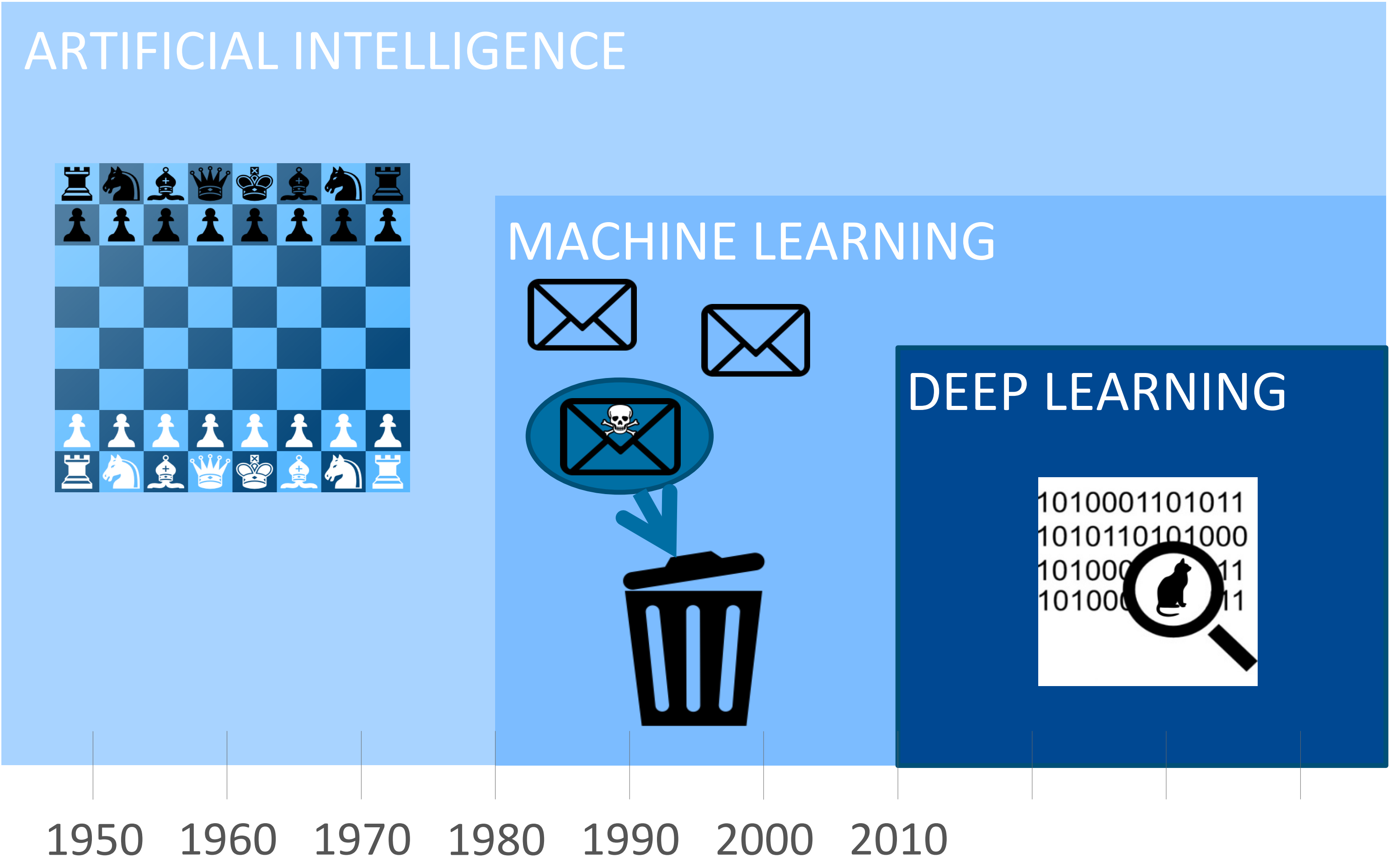
EUROSHNET

ARTIFICIAL INTELLIGENCE
MEETS SAFETY AND HEALTH AT WORK

Trustworthy Artificial Intelligence

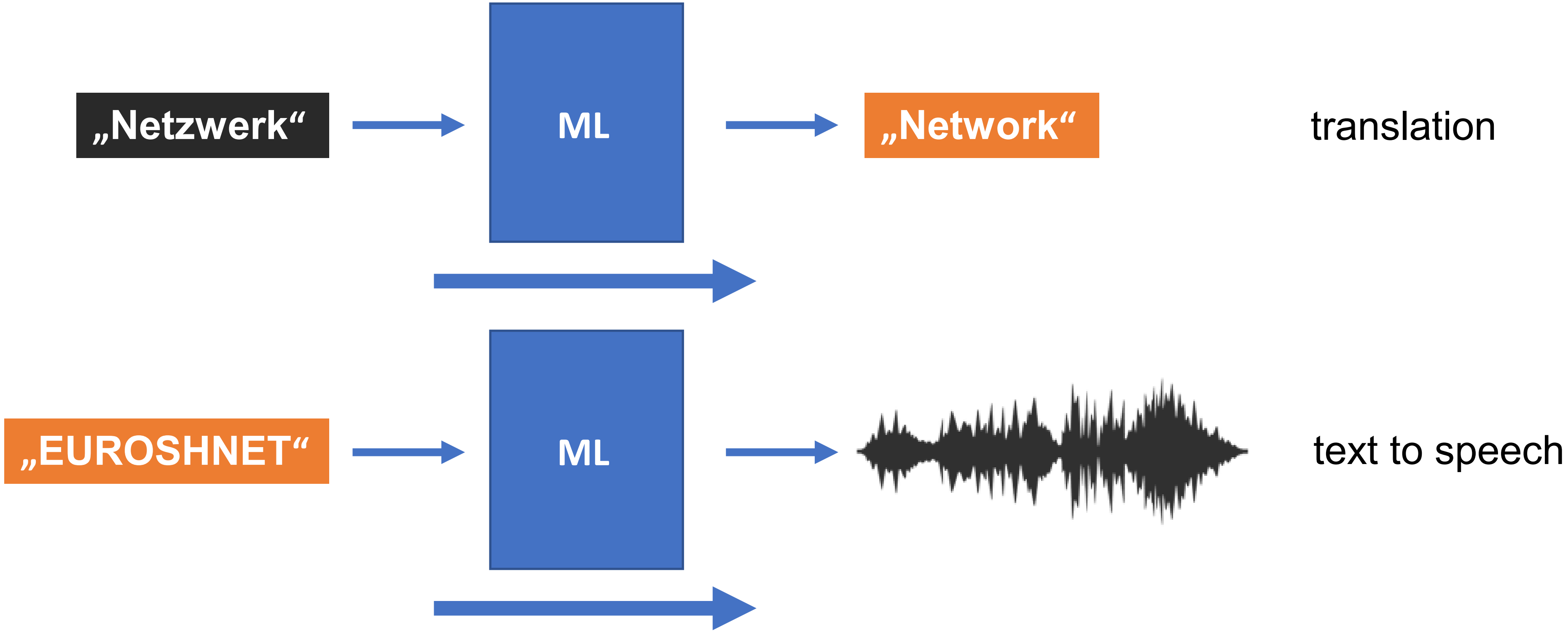
Prof. Dr. André Steimers
Koblenz University of Applied Sciences
Institute for Occupational Safety and Health of the German Social Accident Insurance (IFA)

Overview


ARTIFICIAL INTELLIGENCE
MACHINE LEARNING
DEEP LEARNING
1010001101011
1010110101000
1950
1960
1970
1980
1990
2000
2010

Machine Learning

- In machine learning, a predominantly automated learning process uses sample data to create a model that maps an input to an output



Machine Learning

- In machine learning, a predominantly automated learning process uses sample data to create a model that maps an input to an output

ML

semantic segmentation

ML

„lion“

classification

Examples of some AI errors

- 07/2016 guard robot injures child in department store
- 11/2016 robot Xiao-Pang injures fair visitors
- 05/2017 rear-end collision Tesla model S fire truck
- 01/2018 rear-end collision Tesla model S fire truck
- 05/2018 rear-end collision Tesla model S police vehicle
- 01/2019 collision with oncoming traffic Tesla Model 3
- 08/2019 rear-end collision Tesla Model S tow truck
- 12/2020 malfunction of a service robot in a store

Examples of some AI errors

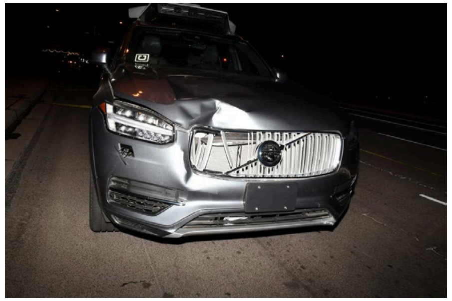
- 01/2016 China, Tesla Model S,1 Driver†
- 05/2016 Florida, Tesla Model S, 1 Driver†
- 03/2018 Arizona, automated Uber Taxi, 1 Pedestrian†
- 03/2018 California, Tesla Model X, 1 Driver†
- 04/2018 Japan, Tesla Model X, 1 Pedestrian†
- 03/2019 Florida, Tesla Model 3, 1 Driver †
- 04/2019 Florida, Tesla Model S, 1 Pedestrian†
- 12/2020 California, Tesla Model S, 2 Persons Honda Civic†
- 05/2020 Norway, Tesla Model X, 1 Pedestrian†

[13]

[14]

[15]

[16]



[17]

[18]

Ethical and safety aspects

- **Trustworthy Artificial Intelligence**
  - Depending on the sources of risk of the selected AI process.

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

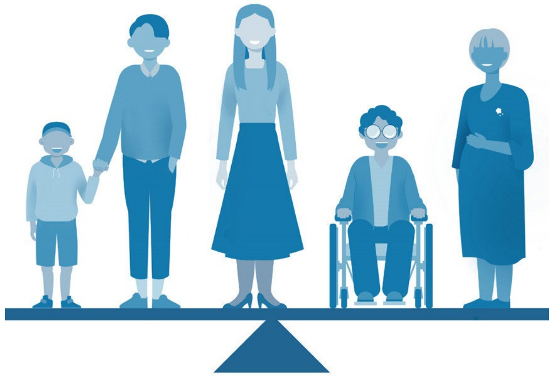
Ethical aspects

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

Fairness

Source: www.jobrapido.com



Source: www.embedica.ai/fairness-2

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

- **Recruiting tool**
  discriminates against women

- **Historic 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

Degree of automation and control

System	Level of automation	Degree of control	Comments
Autonomous	Autonomy	Human out of the loop	The system is capable of modifying its operation domain or its goals without external intervention, control or oversight
Heteronomous	Full automation	Human on the loop Human out of the loop	The system is capable of performing its entire mission without external intervention
	High automation	Human on the loop	The system performs parts of its mission without external intervention
	Conditional automation	Human on the loop	Sustained and specific performance by a system, with an external agent ready to take over when necessary
	Partial automation	Human in the loop	Some sub-functions of the system are fully automated while the system remains und the control of an external agent
	Assistance	Human in the loop	The system assists and operator
	No automation	Human in the loop	The operator fully controls the system

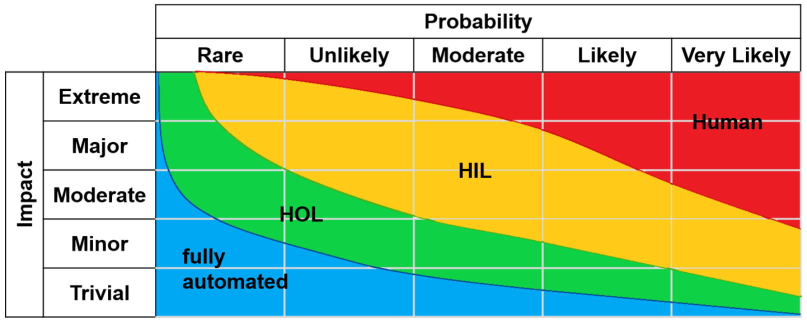
Degree of automation and control

Degree of automation of the system

System

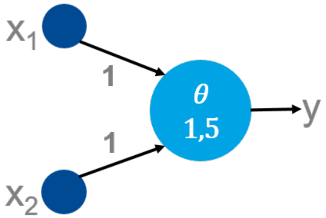
intended use
operational environment
ethical context

Autonomy of the human

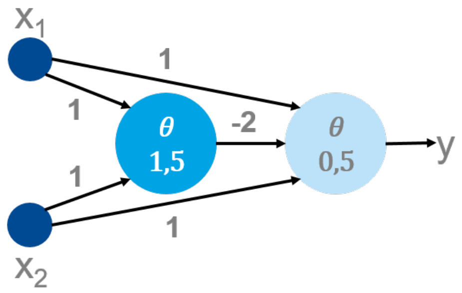


Degree of transparency and explainability

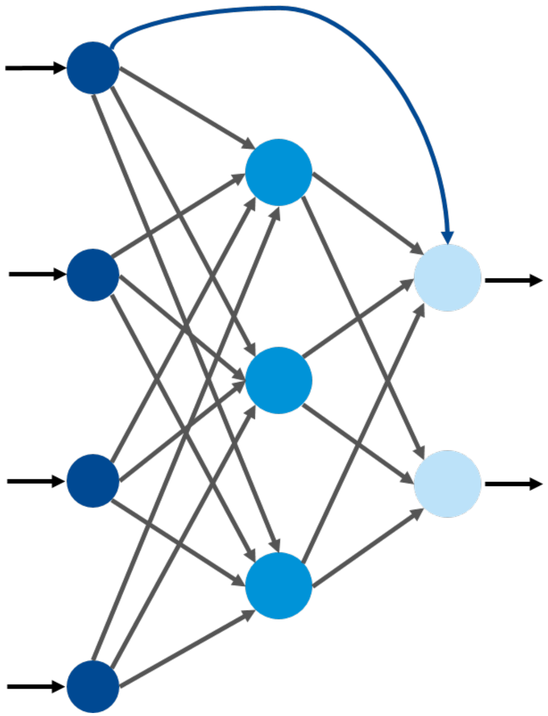
Degree of transparency and explainability



1 Neuron

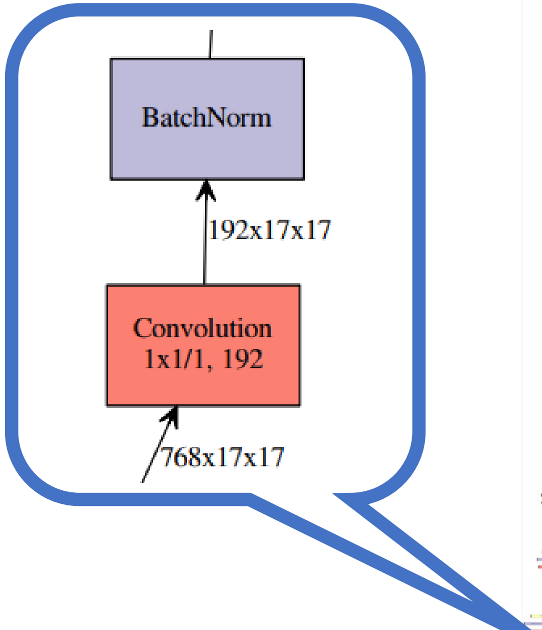


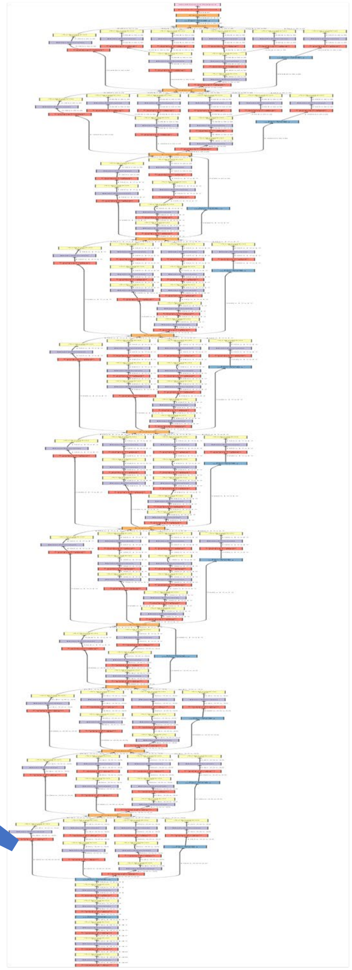
2 Neurons



5 Neurons

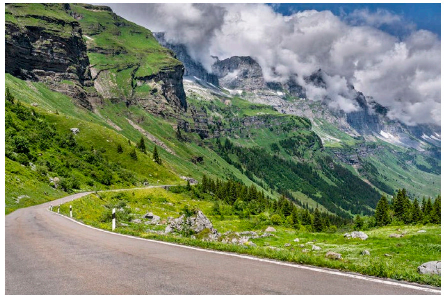
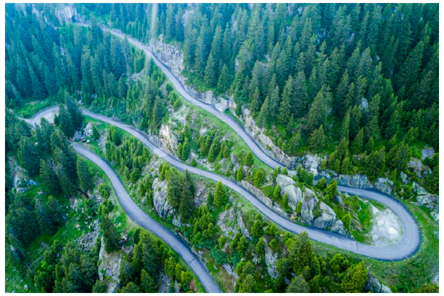
Inception Net V3





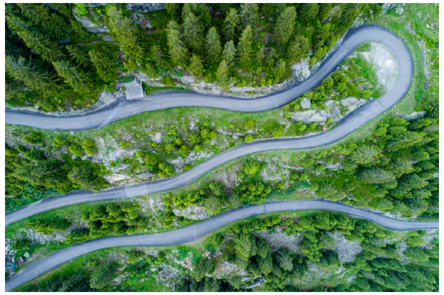
192x17x17 weights
768x17x17 weights

Complexity of the task and usage environment






Complexity of the task and usage environment

- **Completed learning**
  - the model is static and can be extensively validated

- **Continuous learning**
  - the model can adapt to changing environmental conditions

Source: http://georg-dahlhoff.de

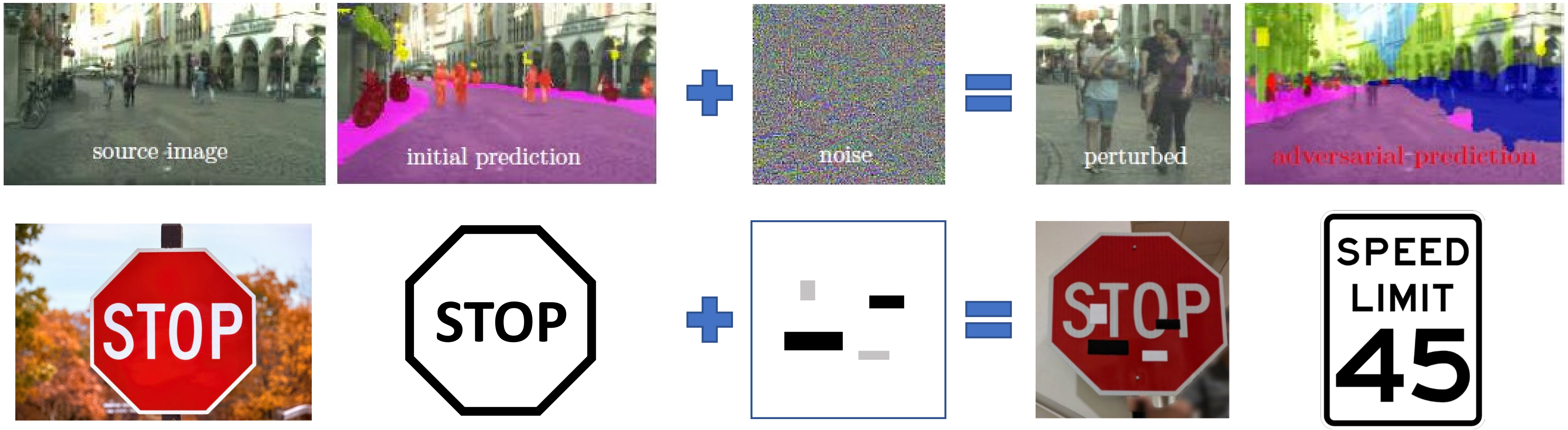
- **Concept Drift**
  - the environment or task of the system deviates from the specification
  - ➢ the system fails because it does not adapt to the new conditions

- **Data Drift**
  - the model differs from the original specification
  - ➢ no static version exists that could be validated

Security

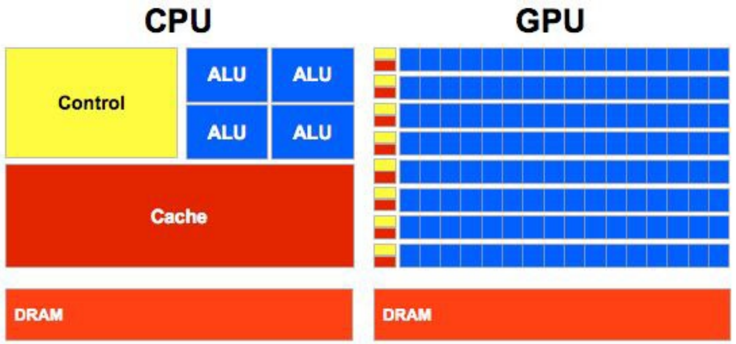
Adverserial Attacks

- A valid model is supplied with disturbed input data 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

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

➢ Translation errors

Source: www.nvidia.com

Thank you for your attention

International Journal of
Environmental Research and Public Health

MDPI

Article

Sources of Risk of AI Systems

André Steimers * and Moritz Schneider

Institute for Occupational Safety and Health of the German Social Accident Health Insurance (IFA), 53757 Sankt Augustin, Germany; moritz.schneider@dguv.de
* Correspondence: andre.steimers@dguv.de

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.

Prof. Dr. André Steimers
Koblenz University of Applied Sciences
steimers@hs-koblenz.de

More Information:

Steimers A, Schneider M. **Sources of Risk of AI Systems**. Int J Environ Res Public Health. 2022 Mar 18;19(6):3641. doi: 10.3390/ijerph19063641