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)

Machine Learning

In machine learning, a predominantly automated learning process uses sample data to create a model that maps an input to an output (e.g. translation, text to speech, semantic segmentation, 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
- 01/2016 China, Tesla Model S, 1 Driver dead
- 05/2016 Florida, Tesla Model S, 1 Driver dead
- 03/2018 Arizona, automated Uber Taxi, 1 Pedestrian dead
- 03/2018 California, Tesla Model X, 1 Driver dead
- 04/2018 Japan, Tesla Model X, 1 Pedestrian dead
- 03/2019 Florida, Tesla Model 3, 1 Driver dead
- 04/2019 Florida, Tesla Model S, 1 Pedestrian dead
- 12/2020 California, Tesla Model S, 2 Persons in Honda Civic dead
- 05/2020 Norway, Tesla Model X, 1 Pedestrian dead

Ethical and safety aspects

Trustworthy Artificial Intelligence

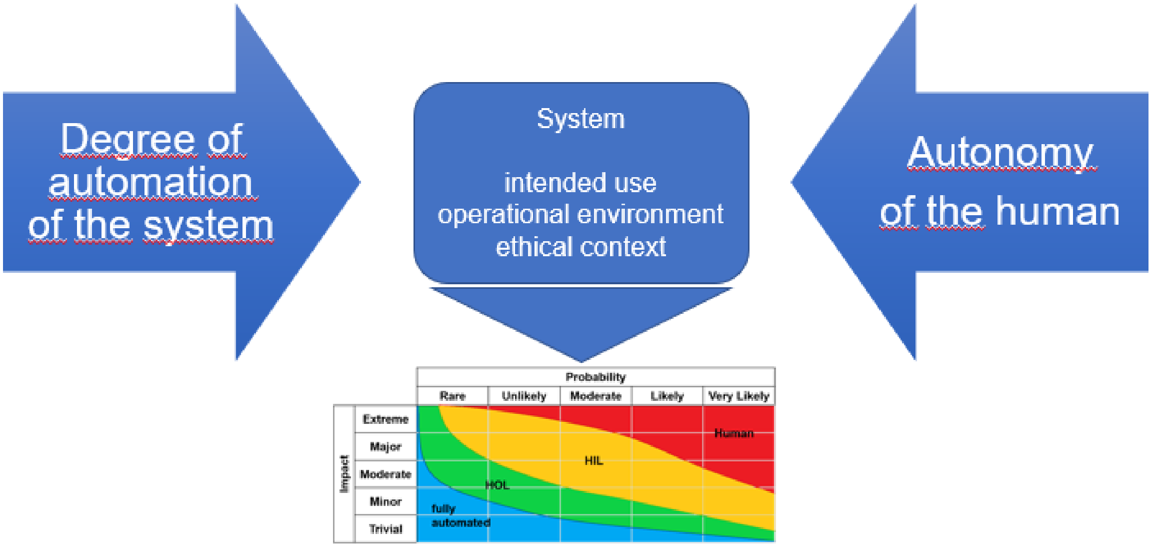
> Depending on the sources of risk of the selected AI process.
>
> Ethical aspects:
>
> 1. Fairness
> 2. Privacy
> 3. Degree of automation and control
>
> Reliability and robustness:
>
> 4. Complexity of the task and usage environment
> 5. Degree of transparency and explainability
> 6. Security
> 7. System hardware
> 8. Technological maturity

Fairness

- *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



Complexity of the task and usage environment

- *Completed learning*
  - the model is static and can be extensively validated
- *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
- *Continuous learning*
  - the model can adapt to changing environmental conditions
- *Data Drift*
  - the model differs from the original specification
  - → no static version exists that could be validated

Security

Adversarial Attacks

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

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

Contact

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