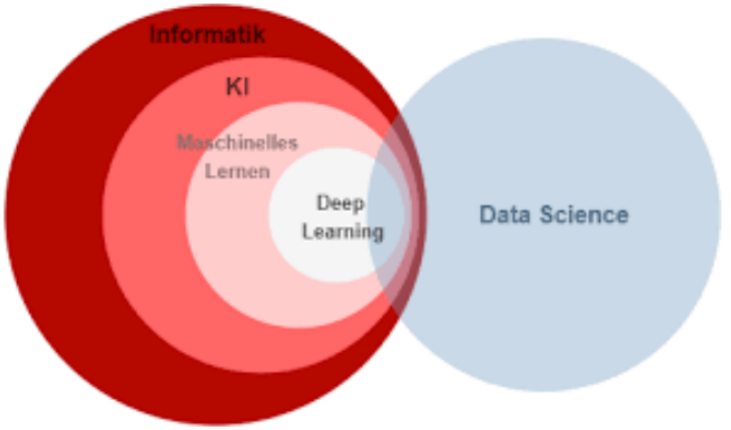
Vertrauenswürdige Künstliche Intelligenz
Anforderungen und Umsetzungsmöglichkeiten

Ute Schmid

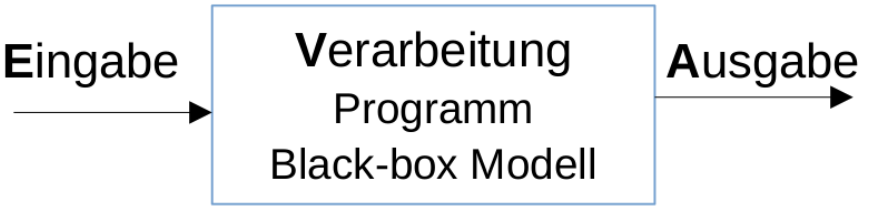
Kognitive Systeme
Otto-Friedrich-Universität Bamberg

Künstliche Intelligenz



- Teilgebiet der Informatik
  - erforscht, wie man Computer dazu befähigt, Dinge zu tun, die Menschen im Moment noch besser lösen können (Rich, 1983)
  - basiert auf der Annahme, dass alle (viele/wichtige) Aspekte menschlicher Intelligenz durch Algorithmen formalisierbar und entsprechend als Computerprogramme simulierbar sind (McCarthy, 1956)
- Digitale Transformation liefert Grundlage für die Anwendung von Algorithmen, auch KI-, speziell ML-Algorithmen

Künstliche Intelligenz

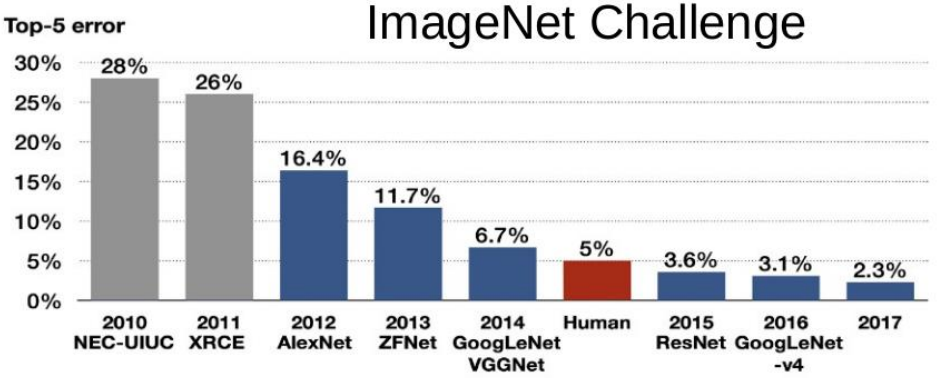


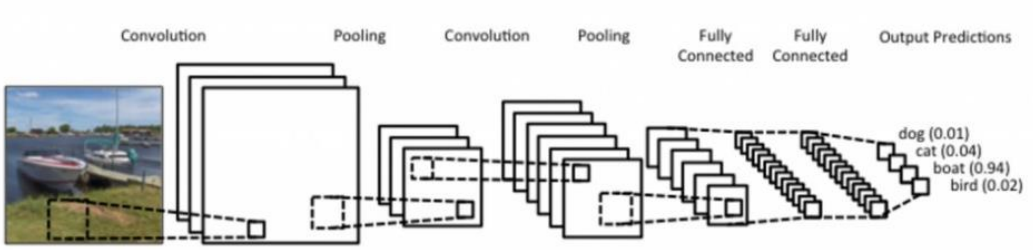
- Die meisten Computerprogramme basieren nicht auf KI-Methoden
- Sobald man KI-Methoden einsetzt gibt man Anforderungen an Korrektheit und Vollständigkeit auf
- KI-Methoden machen Sinn, wenn ein Problem:
  - so komplex ist, dass (optimale) Lösung nicht effizient berechenbar ist → heuristische Methoden, Approximation
  - komplexes (Domänen-) Wissen und Schlussfolgerungen involviert → wissenbasierte Methoden
  - nicht (vollständig) beschreibbar ist → maschinelles Lernen
    Ersetzen von expliziten Algorithmen durch aus Daten gelernte (black-box) Modelle

Deep Learning

- Beeindruckende Erfolge, vor allem bei bildbasierter Klassifikation
- Aber: hoher Aufwand um Daten in ausreichender Anzahl und Qualität zu gewinnen, vor allem in spezialisierten Bereichen

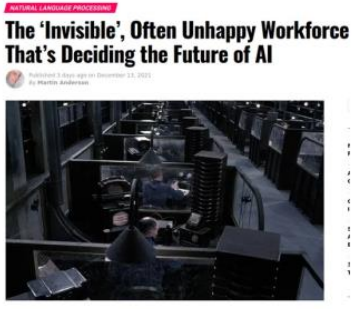
  (*garbage in – garbage out*)



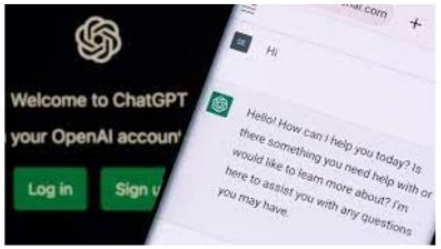


Convolutional Neural Network CNNs (LeCun, 1998)
Alex Krizhevsky, (PhD student of G. Hinton, 2012)

- 4 Mio Bilder
- 1000 Klassen
  - ✗ hand-labelled



Generative KI – ChatGPT



- Von Open AI (sponsored by Microsoft), veröffentlicht 30.11. 2022, Januar 2023: mehr als 100 Mio Nutzende – *fastest growing consumer application to date*
- Basiert auf einer Transformer Network Architektur, die auf der NeurIPS 2017 von Google Brain vorgestellt wurde
- GPT-3: trainiert auf Hunderte von Milliarden von Wörtern, 175 Milliarden Parameter, 800 GB Speicherplatz, 2048 Token langer Kontext
- Es wird geschätzt, dass das Training von GPT-3 1.287 MWh verbraucht und 552 Tonnen CO2 emittiert hat!

Komponenten:

- Sprachmodell: GPT3 – self-supervised learning, alle „crawl“-baren Inhalte bis 2021, dann eingefroren – Übergangswahrscheinlichkeiten zwischen Worten/Tokens
- Überwachtes Lernen: Dialoggenerierung
- Überwachtes Lernen: Filtern toxischer Inhalte
- Human-in-the-loop Reinforcement Learning: Finetuning

Time, 18.Jan 2023

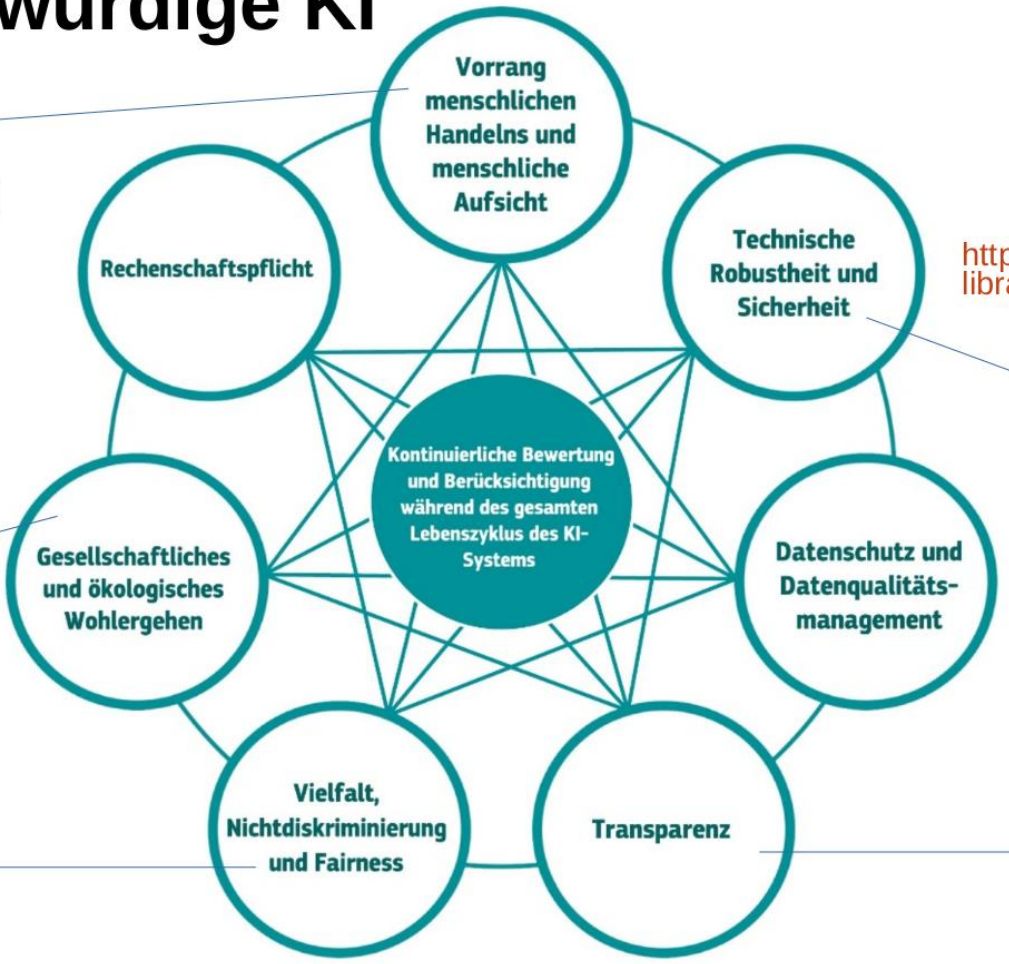
BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

Vertrauenswürdige KI

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Partnerschaftliche KI
Interaktives ML



Hybride KI

Safety & Security

Sozio-technische Einbettung

FairAI

XAI-Methoden

Transparenz – Erklärbare KI (XAI)

- Erklärungen sind nicht pauschal hilfreich
  → Wem erkläre ich was für welchen Informationsbedarf
- **Für Modellentwickler:** overfitting, biases
- **Für Domänen-Expertinnen & -experten:** Nachvollziehbarkeit von Entscheidungen für kalibriertes Vertrauen und *explain to revise*
- **Für Endnutzende:** Transparenz von datenbasierten Entscheidungen (recommender, Kreditvergabe, …)

Volume 36, Issue 3–4

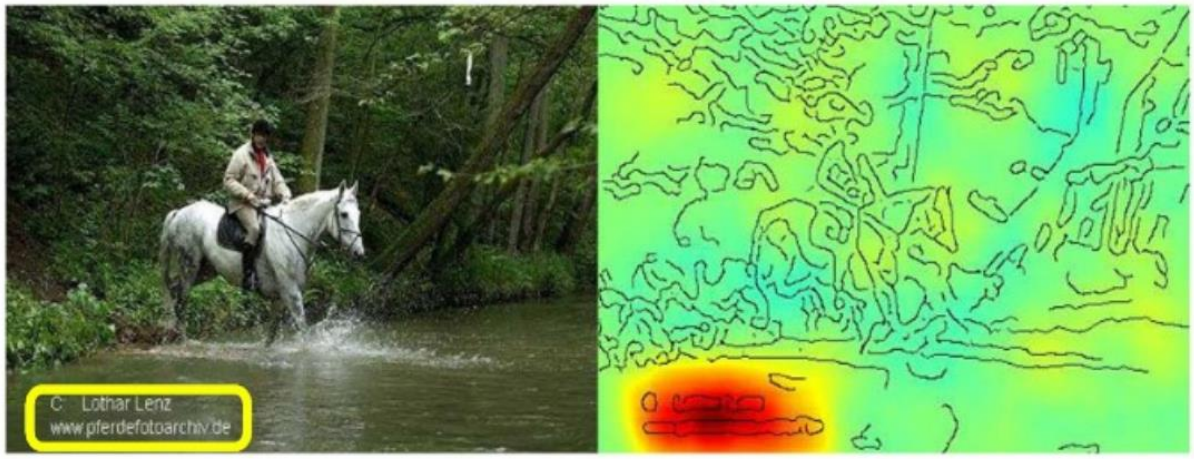
December 2022

Explainable AI

Issue Editors: Ute Schmid, Britta Wrede

XAI kann helfen
Overfitting
zu erkennen

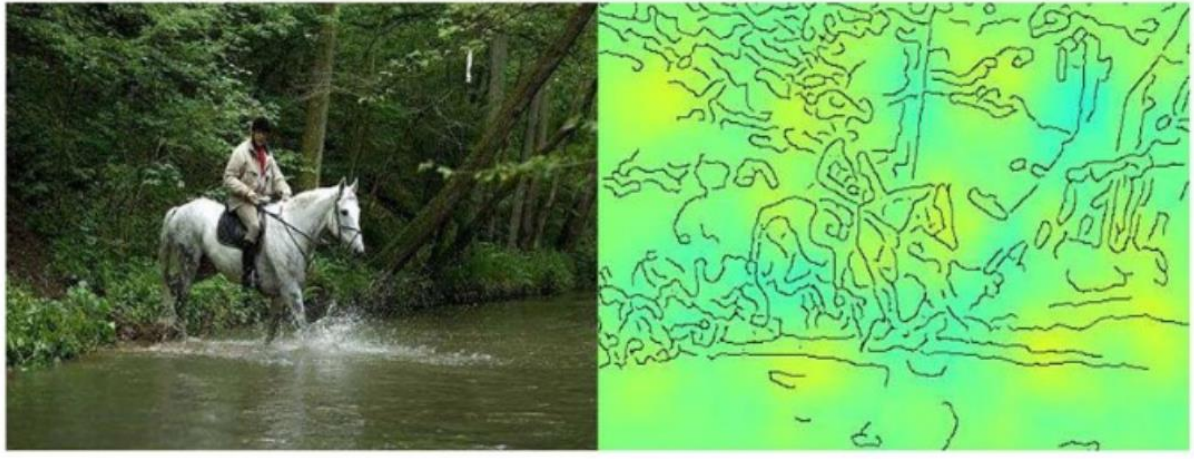
Horse-picture from Pascal VOC data set



Source tag present

↓

Classified as horse



No source tag present

↓

Not classified as horse

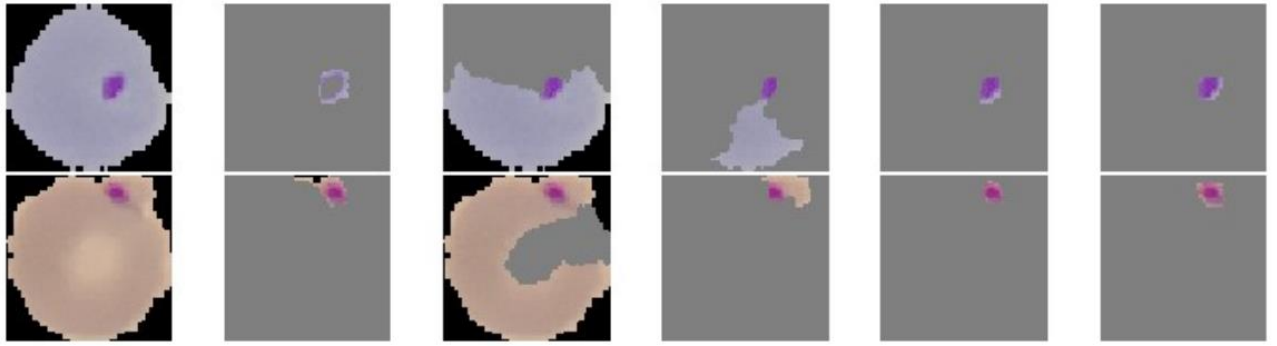
Lapuschkin, Sebastian, et al. "Unmasking Clever Hans predictors and assessing what machines really learn." Nature communications 10.1 (2019): 1096.

Aber: Erklärungen müssen modelltreu sein

Table 2: Jaccard Coeffficient of the different superpixel methods

Superpixel method	Mean Value	Variance	Standard deviation
Felzenszwalb	0.85603243	0.03330687	0.18250170
Quick-Shift	0.52272303	0.04613085	0.21478094
Quick-Shift optimized	0.88820585	0.00307818	0.05548137
SLIC	0.96437629	0.00014387	0.01199452
Compact-Watershed	**0.97850773**	**0.00003847**	**0.00620228**

Schallner, Ludwig, et al. "Effect of superpixel aggregation on explanations in LIME–a case study with biological data." Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I. Springer International Publishing, 2020.



(a) Original (b) Felzen-szwalb (c) Quick-Shift (d) Quick-Shift opt. (e) SLIC (f) Compact-Watershed

Fig. 4: LIME results for true positive predicted malaria infected cells

Explainability 2.0: Konzept-basierte Erklärungen

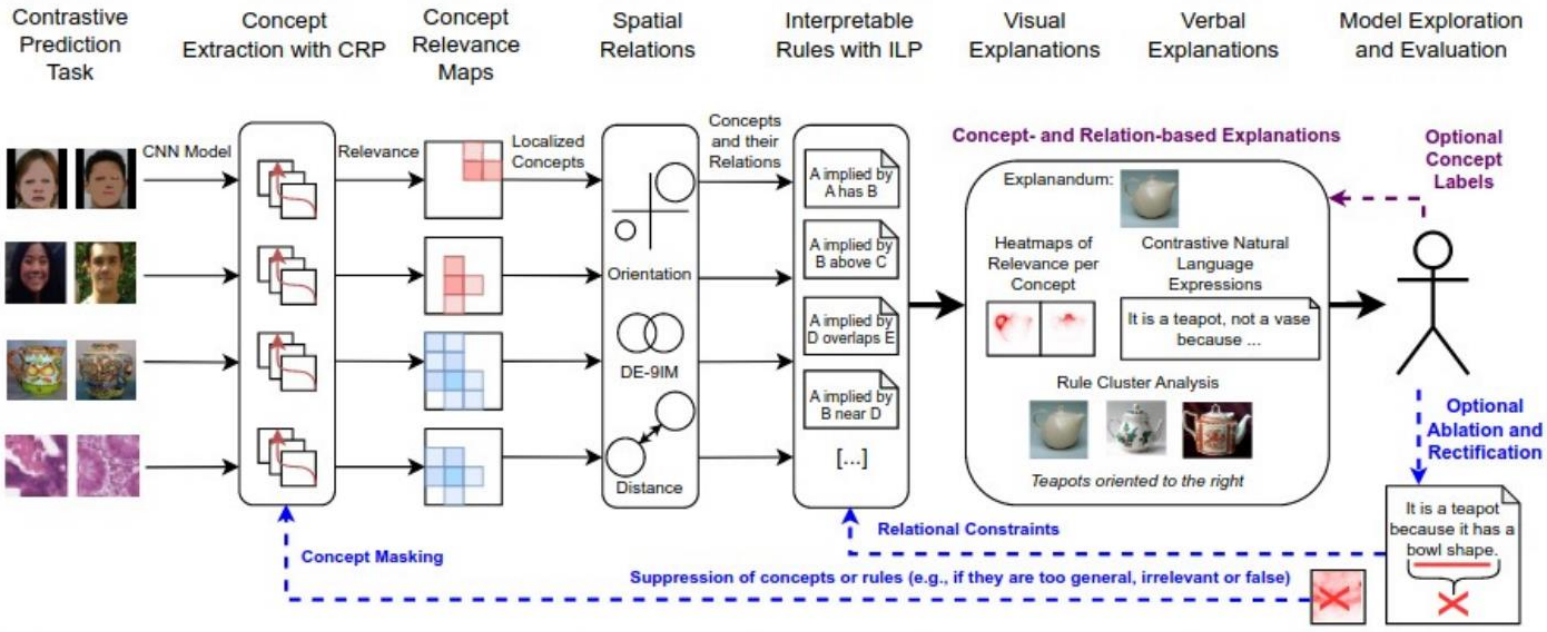
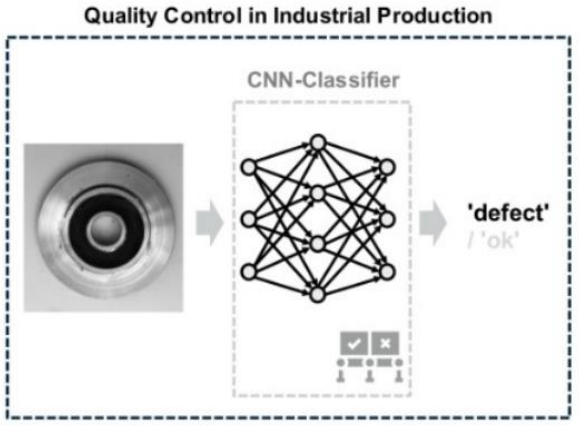
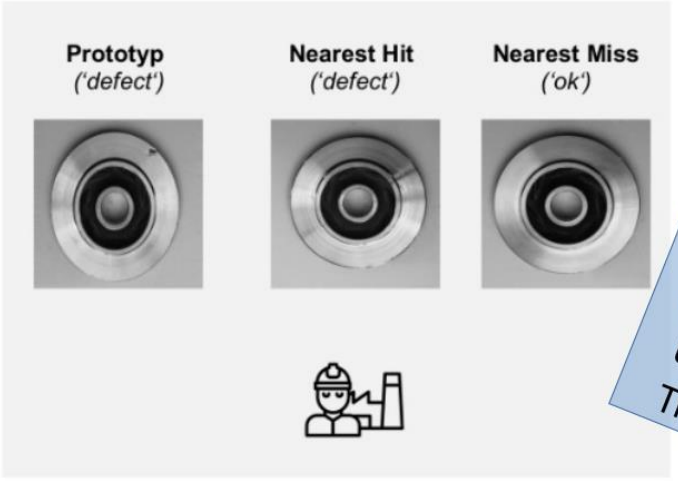


Fig. 2: Overview of our CoReX approach for explaining and evaluating CNN image classifications with concept- and relation-based explanations and constraints (concept masking and relational constraints).

Finzel, Hilme, Rabold, Schmid (u.r.), Rectifiable Concept- and Relation-based Explanations, MLJ

Example-based Explainable AI (XAI) Demonstrator





Unterstützt Domänenexperten, die Entscheidungsgrenzen des gelernten Modells zu verstehen und entsprechend durch neue Trainingsdaten zu korrigieren

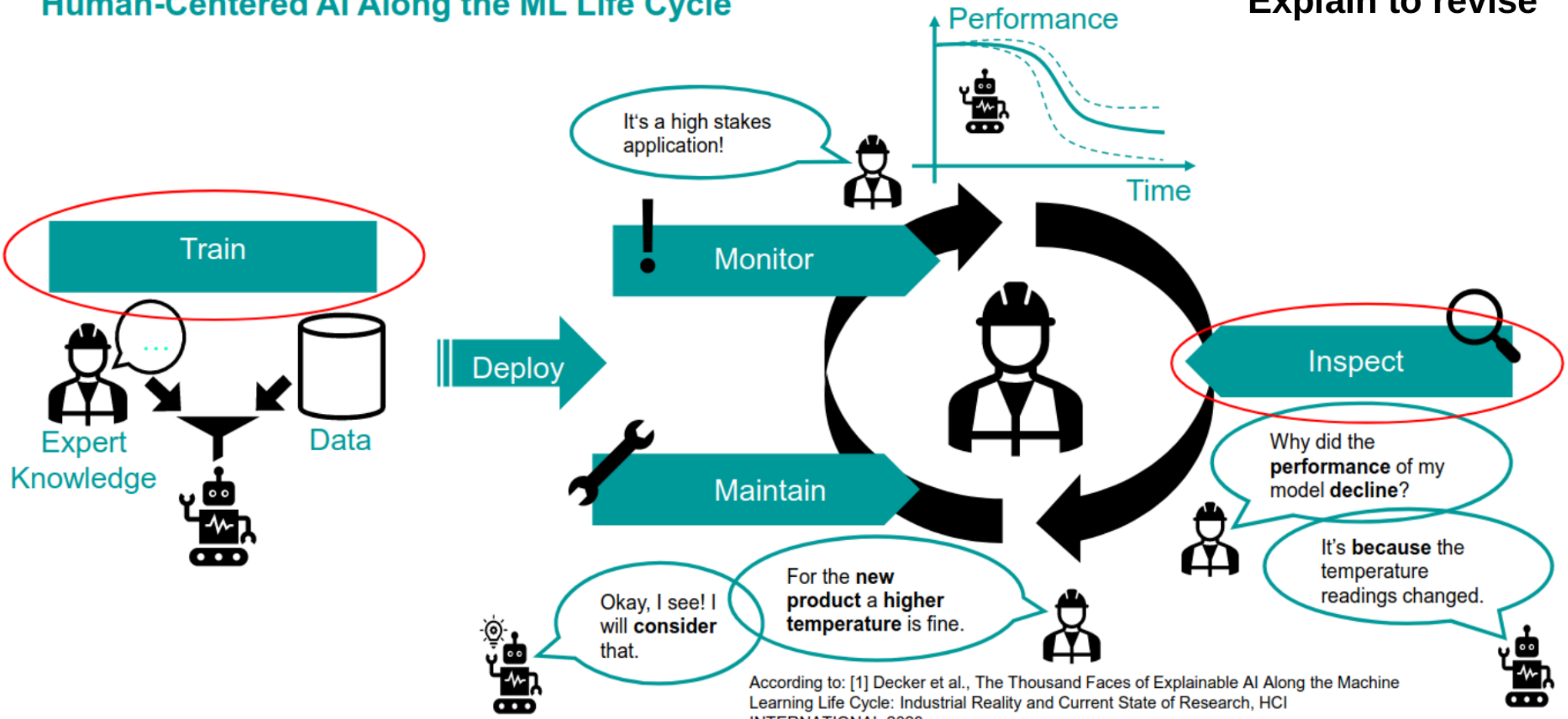
Re-implementation of Kim, Khanna, Koyejo: Examples are not Enough – Learn to Criticize! Criticism for Interpretability, NeurIPS 2016

$$\mathrm{MMD}^2(X,Y) := \frac{1}{|X|^2}\sum_{x_1,x_2\in X} k(x_1,x_2) + \frac{1}{|Y|^2}\sum_{y_1,y_2\in Y} k(y_1,y_2) - \frac{2}{|X|\cdot|Y|}\sum_{x\in X, y\in Y} k(x,y)$$

Herchenbach, Marvin, et al. "Explaining Image Classifications with Near Misses, Near Hits and Prototypes: Supporting Domain Experts in Understanding Decision Boundaries." Pattern Recognition and Artificial Intelligence: Third International Conference, ICPRAI 2022, Paris, France, June 1–3, 2022, Proceedings, Part II. Cham: Springer International Publishing, 2022.

Human-Centered AI Along the ML Life Cycle

Explain to revise



According to: [1] Decker et al., The Thousand Faces of Explainable AI Along the Machine Learning Life Cycle: Industrial Reality and Current State of Research, HCI INTERNATIONAL 2023

SIEMENS

Take Away

Stuart Russell: *We never asked ourselves „what if it really works“ (2019)*

- Aktuelle KI-Methoden (Deep Learning, Generative KI) haben großes Potential für viele Anwendungsfelder (Medizin, Produktion, Verwaltung, Bildung, …)
- Für eine Umsetzung in der Praxis sind Robustheit, Nachvollziehbarkeit und Korrigierbarkeit relevant
- Anforderungen für vertrauenswürdige KI sind Forschungstreiber: Methoden für XAI, Interaktives/human-in-the-loop Lernen, wissensinformiertes ML
- Für Klassifikation gibt es bereits zahlreiche Ansätze für Vertrauenswürdigkeit, für generative KI beginnt die Forschung dazu gerade
- Die Umsetzung des europäischen AI Acts ist jedoch nicht trivial (Zertifizierung)