Trustworthy Artificial Intelligence

Comprehensible, Knowledge-informed, and Revisable

Ute Schmid

Cognitive Systems

Otto-Friedrich-Universität Bamberg

Zentrum für Vertrauenswürdige KI

EU Requirements Trustworthy AI

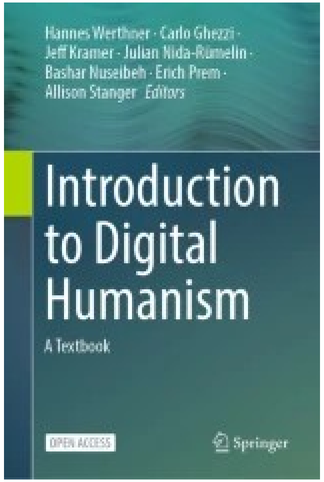
https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

In this Lecture:

- Intro to AI/ML
  - What are the challenges for trustworthiness?
- Selected methods to enhance trustworthiness
  - XAI
  - Hybrid AI
  - Interactive ML

What is your assessment of trustworthiness of current/future AI systems?

- In what application domains would you trust
  - autonomous AI systems?
  - human-supervised AI systems?
- In what application domains would you not trust AI systems?
- Are there specific AI approaches in which you would put more/less trust?



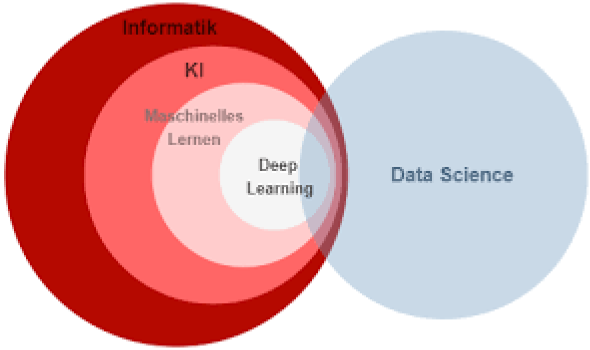
Trustworthy Artificial Intelligence: Comprehensible, Transparent and Correctable

Ute Schmid

Pages 151-164 | Open Access

Part I
General Introduction to AI and ML

Artificial Intelligence (AI)



- 1956 (John McCarthy, Stanford)
- As part of computer science/informatics
  - Based on the assumption that all (many/relevant) aspects of human intelligence can be formalized by algorithms and simulated by computer programs
  - AI is the study of how to make computers perform intelligent tasks that, in the past, could only be performed by humans (Elaine Rich, 1983)
- Digital transformation provides for applicability of algorithms, also of AI algorithms

https://kompetenzzentrum-hamburg.digital/digitaler-glossar/kuenstliche-intelligenz

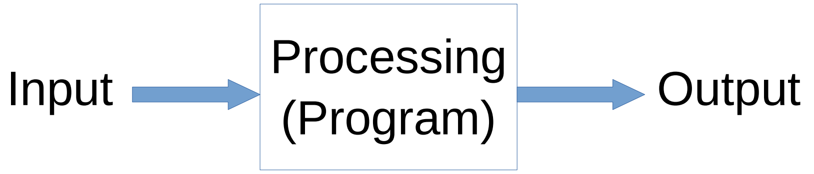
AI vs Standard Computer Science

- Most computer programs are not based on AI methods!
- Application of AI methods means to give up requirements concerning **correctness** and **completeness**

  e.g. air bag controller needs guarantees that it opens by impact and does not open in other situations, and this must hold for all possible situations
- Standard programs can be inspected, systematically tested, it can be proved that core requirements are fulfilled
- However, complex standard software also can have errors

Need for AI Methods

- AI methods are applied if:
  - A problem is so complex that its (optimal) solution cannot be computed efficiently → heuristic methods, approximation
  - A problem involves complex (domain) knowledge and requests valid inferences → knowledge based methods
  - A problem cannot be desrcibed explictly → machine learning, replacement of explicit algorithms by (**blackbox**) models induced from data



Three Waves of AI

- **1. Wave: Focus on explicit knowledge representation**
  - Powerful inference methods, provable characteristics, transparent/comprehensible
  - Expert Systems
  - But: Polanyi's Paradox – *How can we know more than we can tell?*
  - Large amount of knowledge is tacit, implicit, not verbalizable

Great expectations – big disappointments

→ **AI Winter**

1974-1980: only toy problems

1987-1993: Knowledge Engineering Bottleneck

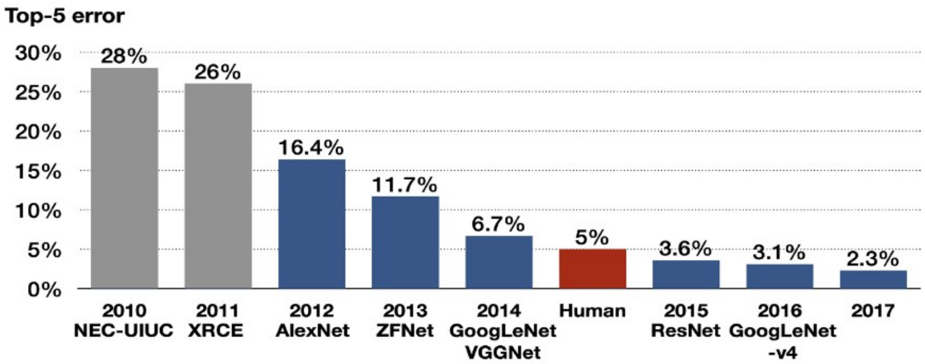
2000-2008: „Winter without end"

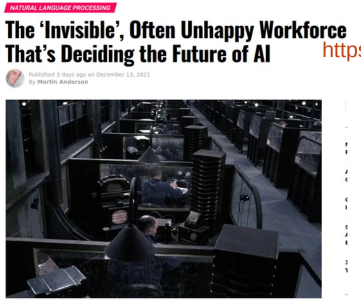
==> **Big Bang of Deep Learning**

Three Waves of AI

2. Wave: Focus on machine learning

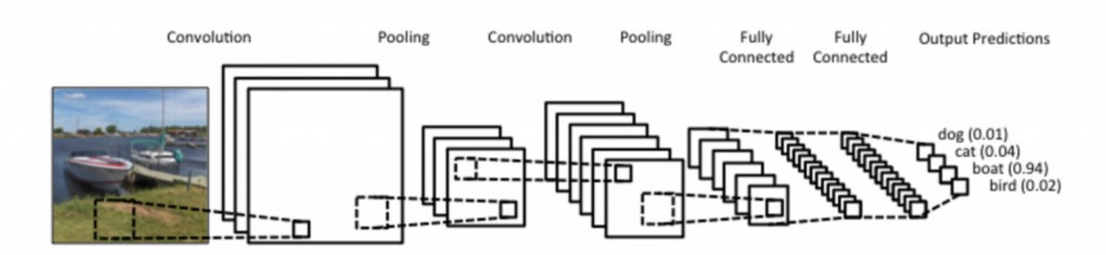
- Impressive successes, especially for image-based classification (end-to-end learning)
- Hope: Replace thinking about a problem by sampling data
- But: high effort to obtain data in sufficient quantity and quality, especially in specialized areas (*garbage in – garbage out*)





https://www.implantology.or.kr/articles/pdf/RvNO/kaomi-2020-024-03-5.pdf

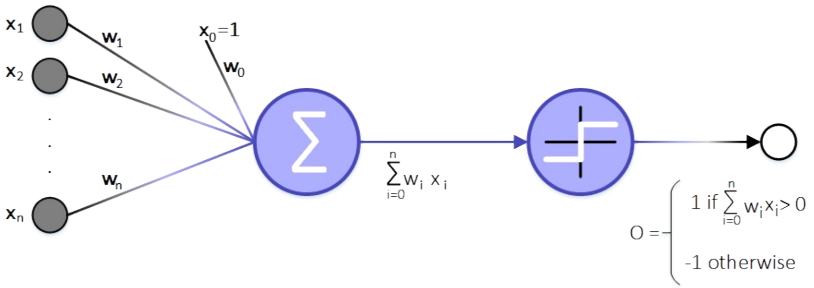
ImageNet Challenge:
4 Mio images,1000 categories, annotated by humans

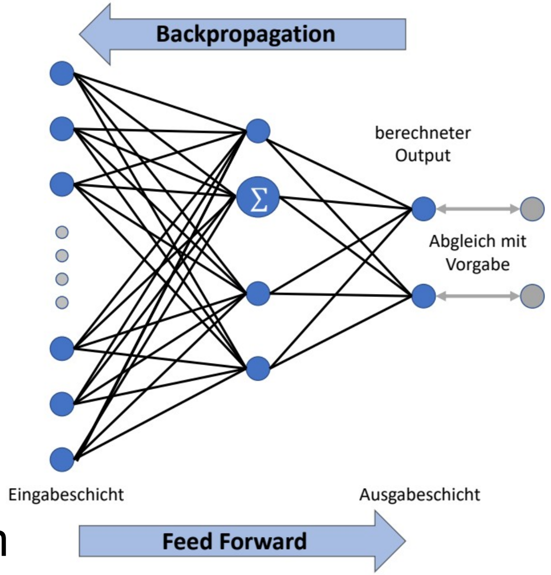


Convolutional Neural Network CNNs (LeCun, 1998)
Alex Krizhevsky, (PhD student of G. Hinton, 2012)

Machine Learning (ML)

- ML is more than neural networks
- Perceptron (Rosenblatt, 1958)
- Reinforcement Learning (Michie, 1961; Sutton, 1998)
- Feed Fordward Neural Networks Backpropagation (LeCun, Rumelhart, Hinton, since 1975)
- Decision Tree Learning (Quinlan, 1985)
- Inductive Logic Programming (Muggleton, 1991)
- Support Vector Machines (Vapnik, 1995) → statistical ML
- AdaBoost (Freund & Schapire,1995), Random Forests (Breiman, 2001),
- Recurrent Networks → Long Shorterm Memory LSTM (Sepp Hochreiter & Jürgen Schmidhuber 1997)





- Learning = Adaptation of weights to optimize performance (wrt to loss)
- Neuronal Networks are blackboxes

HABA Education Neuronal Networks

From Perceptrons to Deep Learning

- We focus on classification learning (supervised, other approaches: generative, representation learning)

  Given a sample of re-labeld training data

  learn a function f: X → Y (binary = concept learning, metric = regression learning)
- Perceptron: adapt weights by simple methods (e.g., just add/subtract input values)
- Multi-Layer-Perceptrons: can learn arbitrary computable functions up to some error (given enough training data and time)
- Convolutional Neural Networks:
  - Learn from raw data (e.g. bitmaps)
  - No need for pre-processing (feature extraction): learn filters together with classification

Sampling Biases

e.g. gender bias
Amazon Recruiting Tool
2015
Rating applicants for
software developer jobs

e.g. ethnic bias
Google Photos

PetaPixel News Reviews Guides Learn Equipment Glossary

Google's Photos App is Still Unable to Find Gorillas

MAY 22, 2023 PESALA BANDARA

Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars

Racial bias in a medical algorithm favors white patients over sicker black patients

AI expert calls for end to UK use of 'racially biased' algorithms

AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.

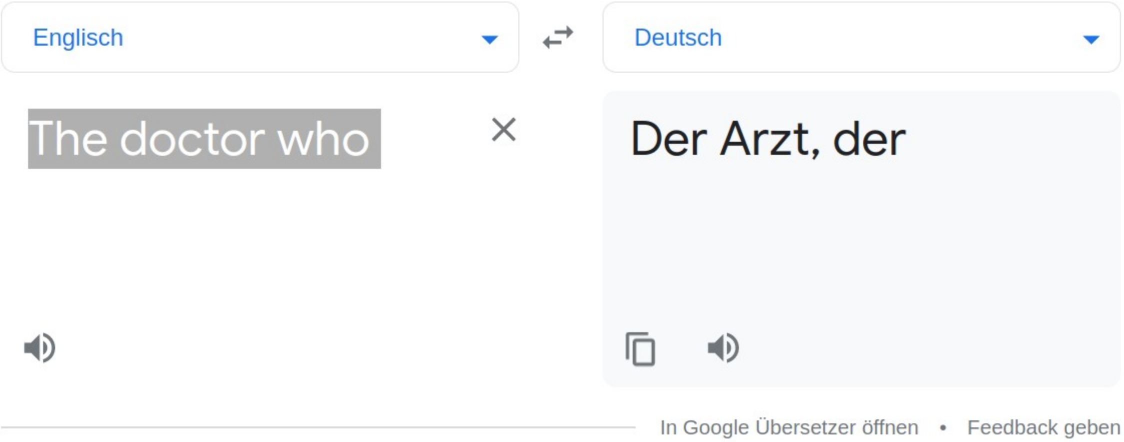
Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

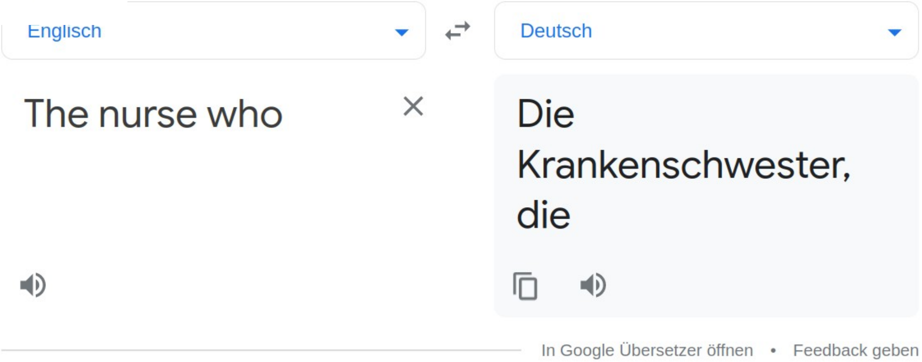
Artificial Intelligence has a gender bias problem – just ask Siri

The Best Algorithms Struggle to Recognize Black Faces Equally

US government tests find even top-performing facial recognition systems misidentify blacks at rates five to 10 times higher than they do whites.

Bias in Machine Translation

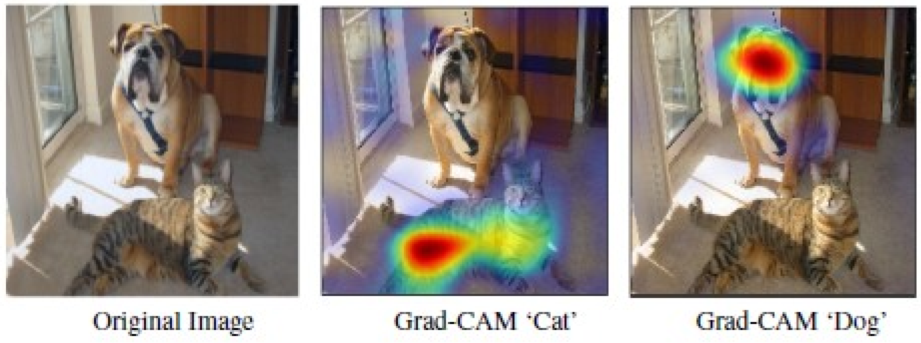




Jonas Troles & Ute Schmid (WMT 2021). Extending Challenge Sets to Uncover Gender Bias in Machine Translation – Impact of Stereotypical Verbs and Adjectives

Three Waves of AI

- **3. Wave: Explainable AI (XAI)**
  - Need for transparency/comprehensibility
  - New family of approaches (starting 2016, see part II):
    - feature relevance (saliency)
    - Concept-based
    - Example-based



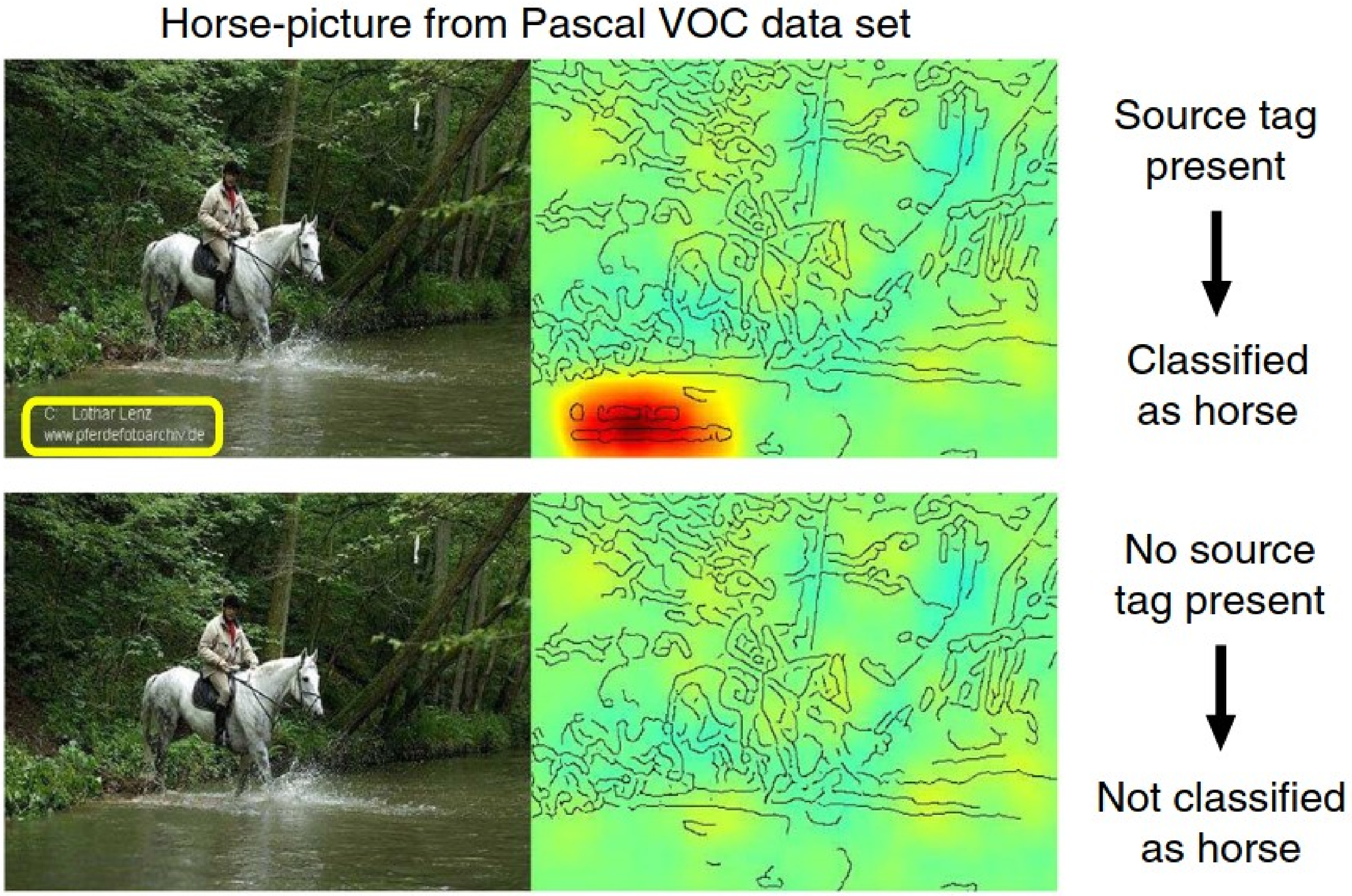
Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization." International Journal of Computer Vision 2019.

  - Soon extended to:
    - hybrid AI/neuro-symbolic AI
    - Interactive ML
    - trustworthy AI

Clever Hans Models

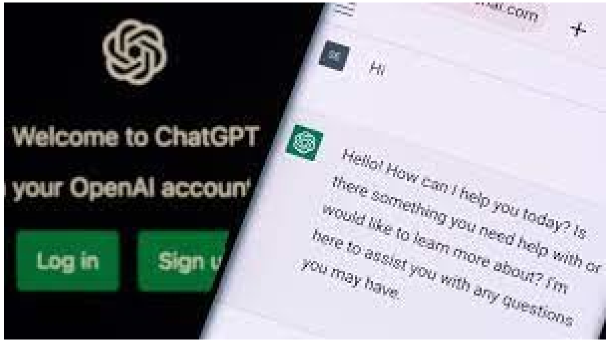
Becoming Aware of Overfitting

i.e. correlation of Irrelevant features With class prediction



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

Generative AI -- ChatGP



- Transformer Networks (Google, 2017)
- Large Language Models, BERT (Google, 2018)
- GPT2 (OpenAI, 2019), GPT3 (2020), GPT4 (2023)
- ChatGPT 30.Nov. 2022 (OpenAI, sponsored by Microsoft), January 2023: > 100 Mio users

- GPT-3: trained on hundreds of billions of words, 175 billion parameters, 800 GB of storage, 2048 tokens of context
- It is estimated that training GPT-3 consumed 1,287 MWh and emitted 552 tons of CO2!

Time, Jan 18 2023

BUSINESS • TECHNOLOGY

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

Components:

- LLM GPT3 – self-supervised learning, data from the web: Common Crawl, books, Wikipedia makes 3%
- Supervised Learning: Generation of dialogue, content filters (toxic content)
- Human-in-the-loop Reinforcement Learning: Finetuning

ChatGPT Reception: From Completely Amazed to Sceptical

GARY MARCUS IDEAS DEC 29, 2022 7:00 AM

WIRED

The Dark Risk of Large Language Models

AI is better at fooling humans than ever—and the consequences will be serious.

TECHNOLOGY

Google Vice President Warns That AI Chatbots Are Hallucinating

This could be a problem.

JAMES FELTON

Senior Staff Writer

Feb 15, 2023 11:53 AM

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

BY BILLY PERRIGO

Time

JANUARY 18, 2023 7:00 AM EST

Content warning: this story contains descriptions of sexual abuse

Training ChatGPT as well as Bard is heavily relying on human feedback

ChatGPT

Is very eloquent and can present wrong information very convincingly

Dec. 19 2022

UT Was ist schwerer ein Kilo Watte oder ein Kilo Stahl?

Ein Kilo Watte ist leichter als ein Kilo Stahl. Stahl ist ein viel dichteres Material als Watte und hat daher eine höhere Masse pro Volumeneinheit. Ein Kilogramm Stahl hat etwa das gleiche Volumen wie ein Würfel mit etwa 12,5 cm Kantenlänge, während ein Kilogramm Watte ein viel größeres Volumen haben würde, da es weniger dichte ist. Die Masse von Stoffen wird in Kilogramm gemessen, während das Volumen in Kubikmetern angegeben wird. Ein Kubikmeter Stahl hat daher eine größere Masse als ein Kubikmeter Watte, da Stahl eine höhere Dichte hat.

Correct answer in English in January 2023

ChatGPT

Has no world knowledge and no mathematical knowledge

Jan. 28 2023

Correct answer
In English, June 6

UT Wenn 5 Personen in einen Bus einsteigen und 8 aussteigen, wieviele Passagiere sind dann im Bus?

Wenn 5 Personen in einen Bus einsteigen und 8 aussteigen, sind dann 2 Passagiere im Bus.

Man kann dies berechnen, indem man die Anzahl der Personen, die einsteigen, von der Anzahl der Personen, die aussteigen, subtrahiert: 5 - 8 = -3, d.h. es sind 3 Personen weniger im Bus als zuvor. Daher sind 2 Passagiere im Bus.

If 5 persons board a bus and 8 persons unboard, how many passengers are in the bus?

If 5 persons board a bus and 8 unboard than there are 2 passengers in the bus. This can be calculated by subtracting the number of persons who board from the ones who exit: 5-8 = -3, that is, there are 3 persons less in the bus. That is 2 passengers.

ChatGPT

Cannot perform common sense reasoning

UT Two children, Cloe and Alexander, went for a walk. They both saw a dog and a tree. Alexander also saw a cat and pointed it out to Cloe. She went to pet the cat. Did Cloe see the cat?

No, according to the information given, Cloe did not see the cat. It was Alexander who saw the cat and pointed it out to Cloe. Cloe went to pet the cat based on Alexander's observation, but it doesn't imply that she actually saw the cat herself.

Example taken from G Marcus, Rebooting AI, asked June 6 2023

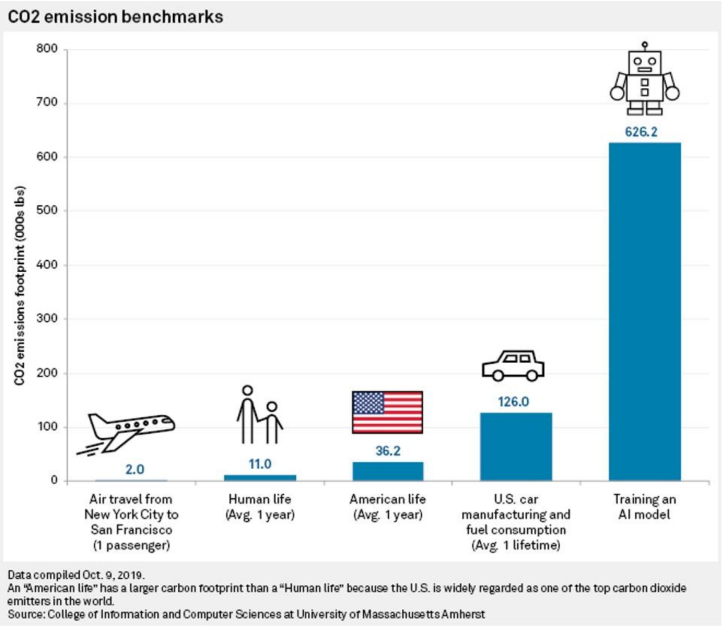
Generative AI

After deep learning the next very powerful approach

- Transformer net (without additional methods):
- has no domain knowledge (e.g. cannot count, see number of fingers in generated images)
- Is trained on data which are not quality controlled and where copyright has not been respected
- Has no direct back relation to data source
- Is highly intransparent
- Might be a stochastic parrot (re-representation of data in transformer, no generalization)

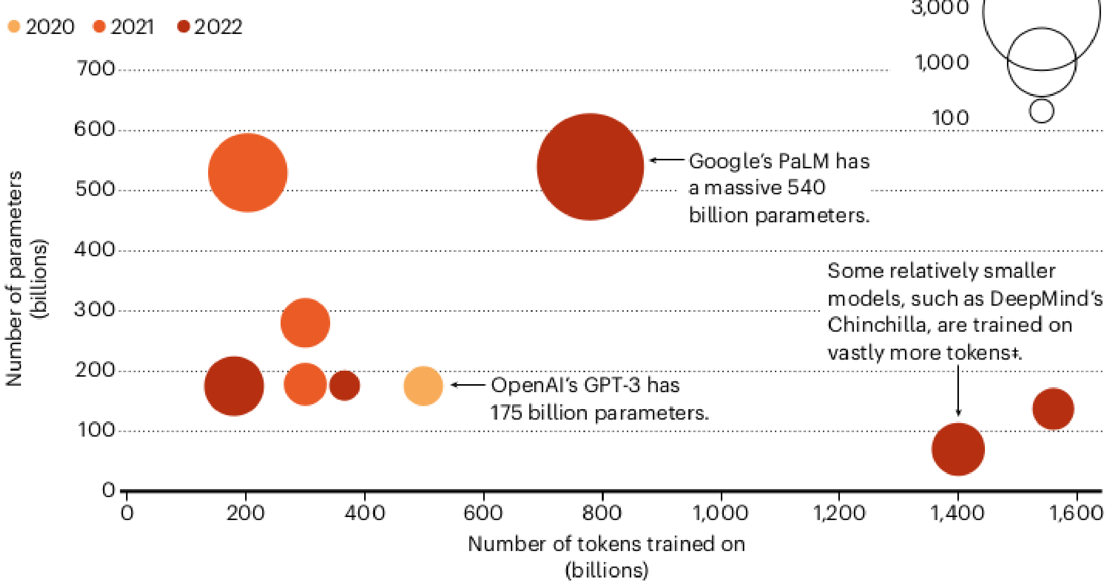
IN AI, IS BIGGER BETTER?

As generative artificial-intelligence models get larger, some scientists advocate for leaner, more energy-efficient systems. **By Anil Ananthaswamy**



DIFFERENT ROUTES TO SCALE

Over the past few years, artificial-intelligence large language models have been trained using more computing power and more parameters*. Some smaller, high-performing models have also appeared, but they are large in another way — they are trained on many more data.



*Parameters: roughly, the number of connections between neurons. †Compute: number of computing operations executed during training, measured as floating point operations (flops). ‡Tokens: words, digits or other units of information that models are trained on.

ADAPTED FROM OUR WORLD IN DATA, AND FROM J. SEVILLA *ET AL.* PREPRINT AT ARXIV HTTPS://DOI.ORG/10.48550/ARXIV.2202.05924 (2022).

ChatGPT vs Search Engine vs Wikipedia

- Search engine: information in the context of a web page – assessment of trustworthiness possible (page of a university clinic, page of a pharmaceutical company, page of a healing stone seller)
- ChatGPT does not allow to refer back to the original source
- Wikipedia: Agile, crowd-sourcing, proven strategy of quality checks by humans

Generative AI

Problems

- ✗ No factual accuracy, no sources (on some topics probably 10% serious content to 90% less serious)
- ✗ Streamlining of language
- ✗ Adoption of US values
- ✗ Copyright
- ✗ Danger of desinformation campagnes
- ✗ Energy demands, CO2 footprint
- ✗ Loss of skills such as structuring complex issues?

Opportunities

- ✔ Relief from more repetitive tasks: more time for understanding, complex problem solving
- ✔ Democratization (writing of text, code generation)

ChatGPT Reception – Synthetic friendliness, `Californication‘

RollingStone MUSIC POLITICS TV & MOVIES (SUB)CULTURE RS RECOMM

UNCANNY VALLEY

Nick Cave Slams AI Attempts at Nick Cave Songs

Fans tasked controversial AI bot ChatGPT to write songs in the musician's trademark style, and he was not amused

BY CHARISMA MADARANG

JANUARY 16, 2023

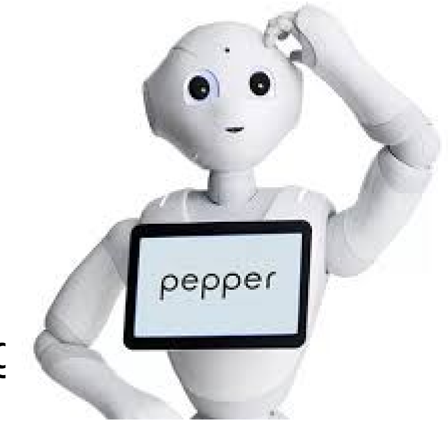
He continued, “Mark, thanks for the song, but with all the love and respect in the world, this song is bullshit, a grotesque mockery of what it is to be human.”

Human Learning

Josh Tenenbaum

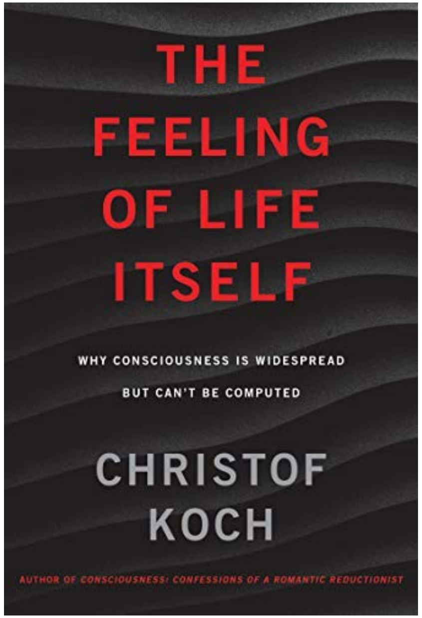
- Inductive Bias (*do not confuse with sampling bias!*)
- Generalization over data is only possible with inductive bias, otherwise one could only store information (rote learning)
- Over-generalization: goed (instead of went)
- Dark side of inductive generalization: Stereotypes and prejudice (girls are not good in math, boys are not good in interpreting poems)

Most AI is not General



- Most AI systems are restricted to one very specific domain
  (weak AI not strong AI)
- A system which is good at classifying animals cannot classify traffic sig
  But: no meta cognition/awareness!
- Inadmissible anthropomorphization!
- Intelligence == excellent chess player, PhD in physics
  vs. building towers from blocks, mixing a drink, recognizing a cat

➔ General AI requires consciousness and intentionality



Summary First Part

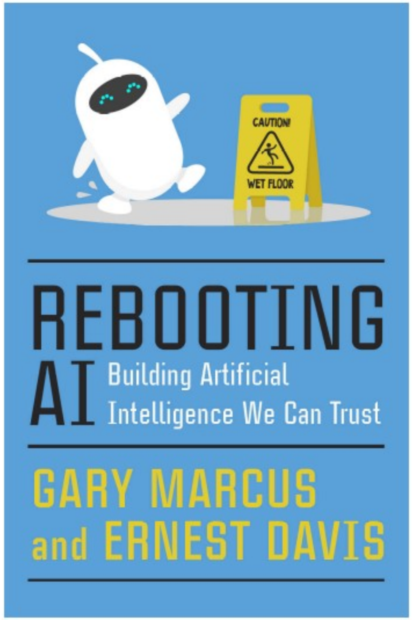
- AI is more than machine learning (knowledge based approaches)
- Learning is inductive generalization over examples
- Supervised learning relies strongly on human input (annotation of ground truth)
- Machine learned models cannot be 100% correct
  - Image search `baby cat on red sofa‘ – what if every 100th image shows something different?
  - Image based medical diagnosis – what if every 100th output is wrong?
- Sampling biases as well as in-equalities in the real world can result in unfair models
- But: machine learning has a lot of merits (if applied adequately)

Part II
Methods for Trustworthy AI

Rebooting AI Reconsidered

(published 2019)

Ernest Davis
University of Bamberg
May 15, 2023



Bottom line

AI has increased much more in power and widespread use than in reliability.

So the problem of building trustworthy AI is not much closer to being solved, but it has become much more URGENT.

Frankfurter Allgemeine Sonntagszeitung vom 19.02.2023

Frankfurter Allgemeine
SONNTAGSZEITUNG

Seite:	53	**Nummer:**	7
Ressort:	Wissenschaft	**Auflage:**	221.832 (gedruckt) [1] 212.008 (verkauft) [1] 220.191 (verbreitet) [1]
Seitentitel:	WISSENSCHAFT	**Reichweite:**	0,740 (in Mio.) [2]
Mediengattung:	Sonntagszeitung		

[1] IVW 1/2022
[2] AGMA ma 2022 Pressemedien II

Viel versprochen

Der Erfolg von ChatGPT war überraschend, seine Grenzen sind es nicht. Wie funktionieren die Sprachbots, wie können sie besser werden - und welche Gefahren bergen sie?

Von Hinnerk Feldwisch-Drentrup

zu jeder Anfrage eine Antwort zu lie-

die Frage "Wer bist du?" wird wahr-

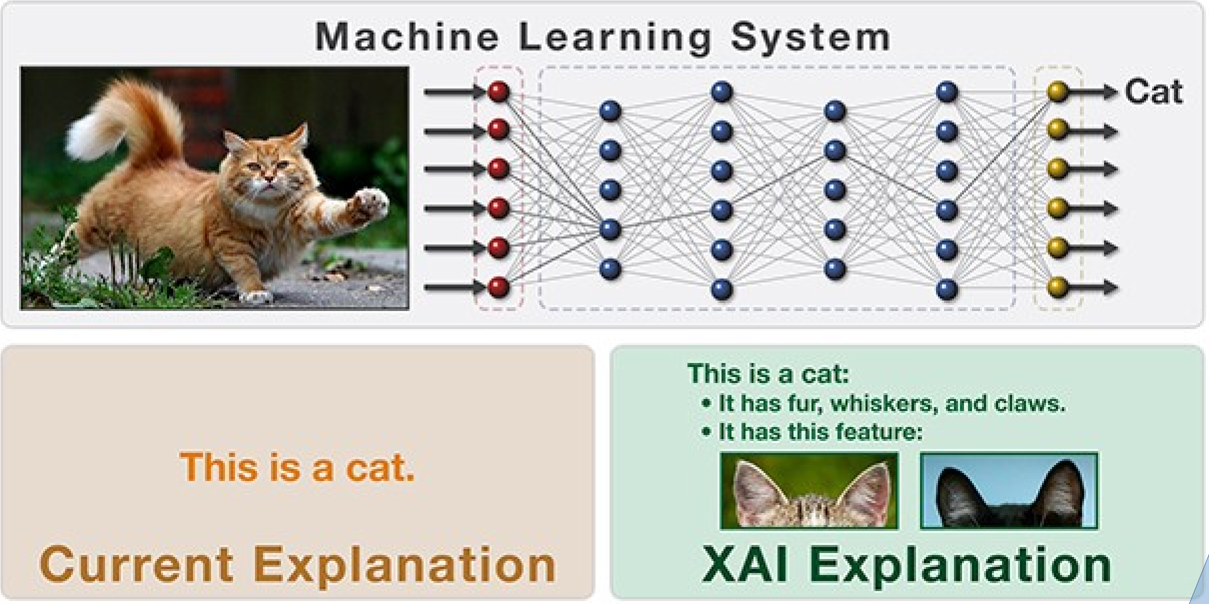
Requirements for Trustworthy AI

European AI Act

Schmid, U. (2024). Trustworthy Artificial Intelligence: Comprehensible, Transparent and Correctable. In In: Werthner, H., et al. Introduction to Digital Humanism. Springer, Cham. https://doi.org/10.1007/978-3-031-45304-5_10

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

David Gunning, IJCAI 2016



Multimodal explanation
Verbal/features
Prototypical examples

http://www.darpa.mil/program/explainable-artificial-intelligence

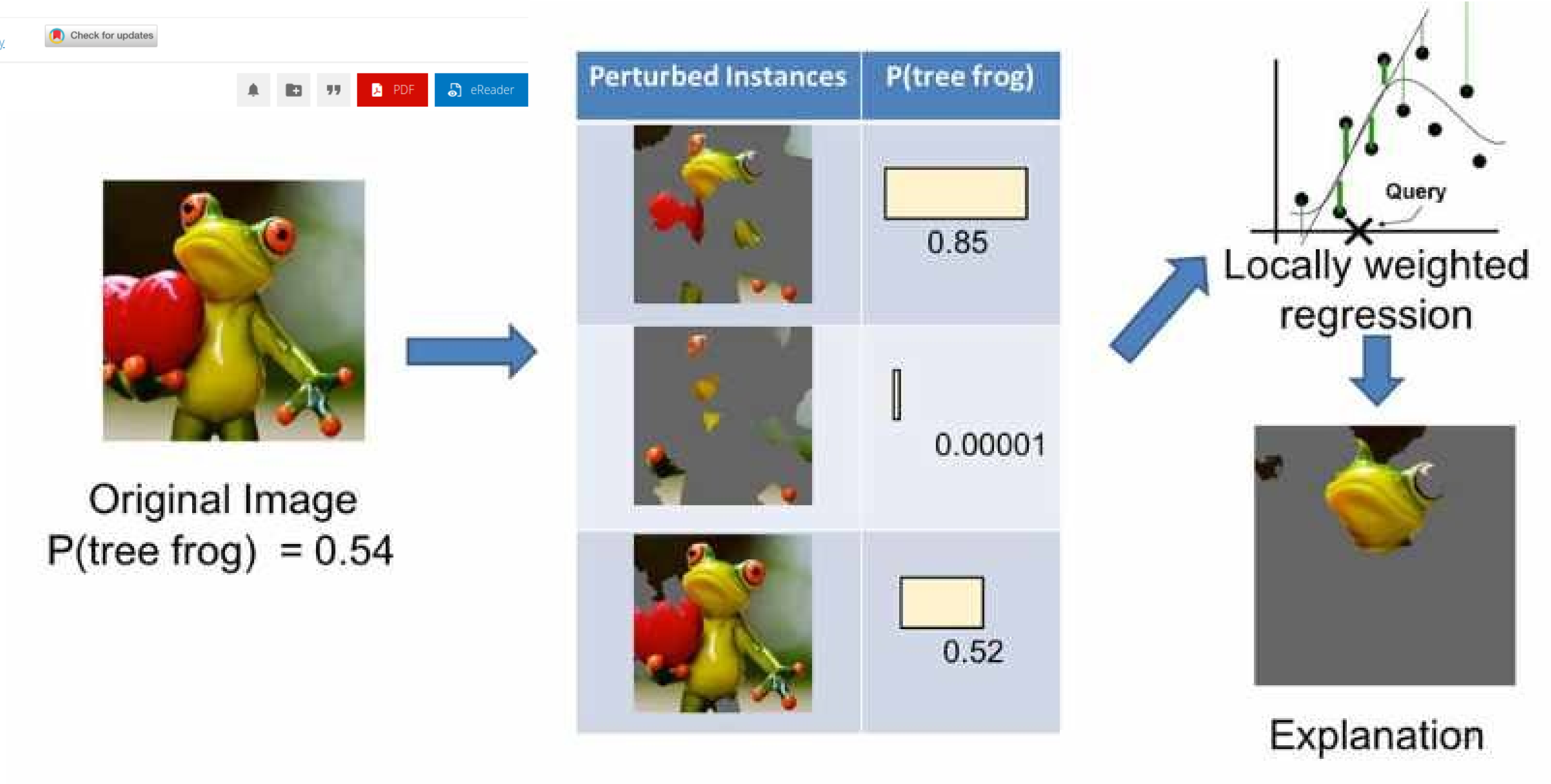
LIME as One of the First XAI Approaches

"Why Should I Trust You?": Explaining the Predictions of Any Classifier

Authors: Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin | Authors Info & Claims

KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining • Pages 1135 - 1144
https://doi.org/10.1145/2939672.2939778

Published: 13 August 2016 Publication History



Transparency – XAI

- Explainability is not useful per se
  → Explain to whom and for what information need

Science

Current Issue First release papers Archive

HOME > SCIENCE > VOL. 373, NO. 6552 > BEWARE EXPLANATIONS FROM AI IN HEALTH CARE

POLICY FORUM | TECHNOLOGY AND REGULATION

Beware explanations from AI in health care

The benefits of explainable artificial intelligence are not what they appear

BORIS BABIC, SARA GERKE, THEODOROS EVGENIOU, AND I. GLENN COHEN Authors Info & Affiliations

SCIENCE · 16 Jul 2021 · Vol 373, Issue 6552 · pp. 284-286 · DOI: 10.1126/science.abg1834

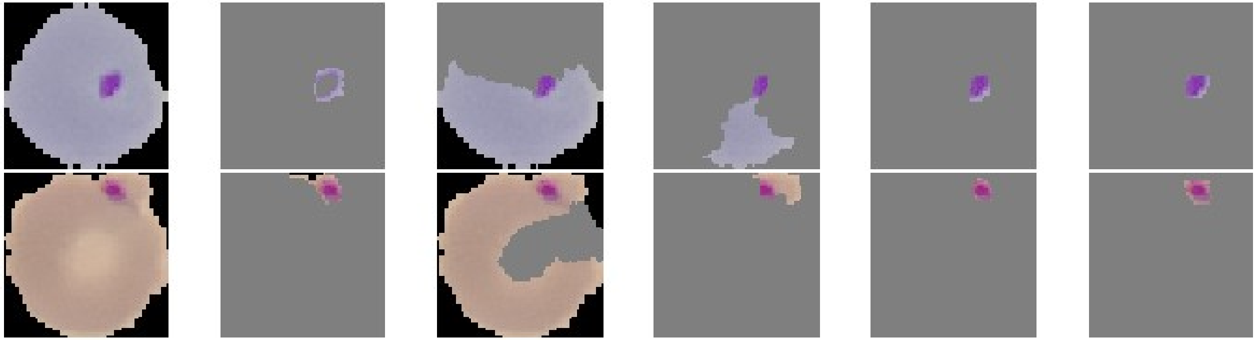
- **For model developers:** overfitting, biases
- **For domain experts: comprehensibility of AI decision making, calibrated (not naive) trust, explain to revise**
- **For end users:** transparency of data-based decision algorithms (insurance, health-apps)

But: Explanations need to be faithful to the model!

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) Felzenszwalb (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: Concept-based Explanations

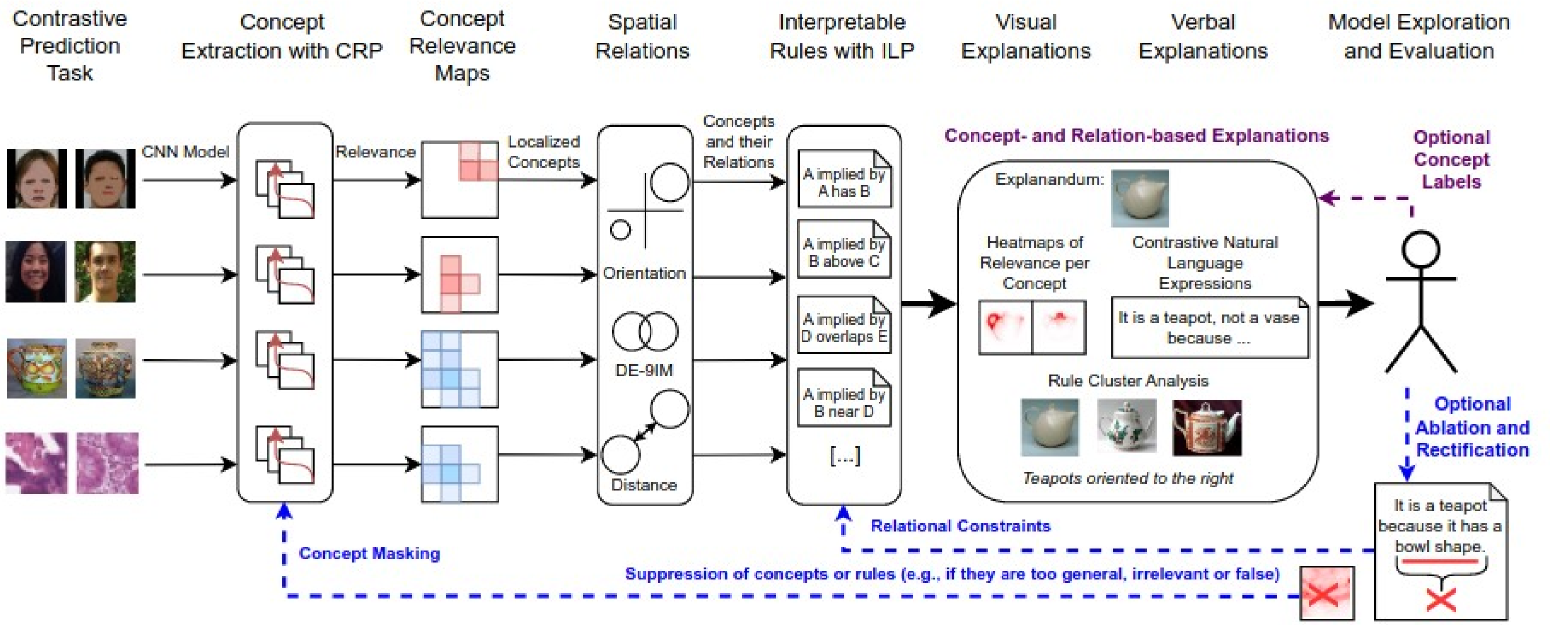


Fig. 2: Overview of our CoReX approach for explaining and evaluating CNN image classifications with concept- and relation-based explanations and constraints (concept masking and relational constraints).

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

Near Miss Explanations for Effective Teaching

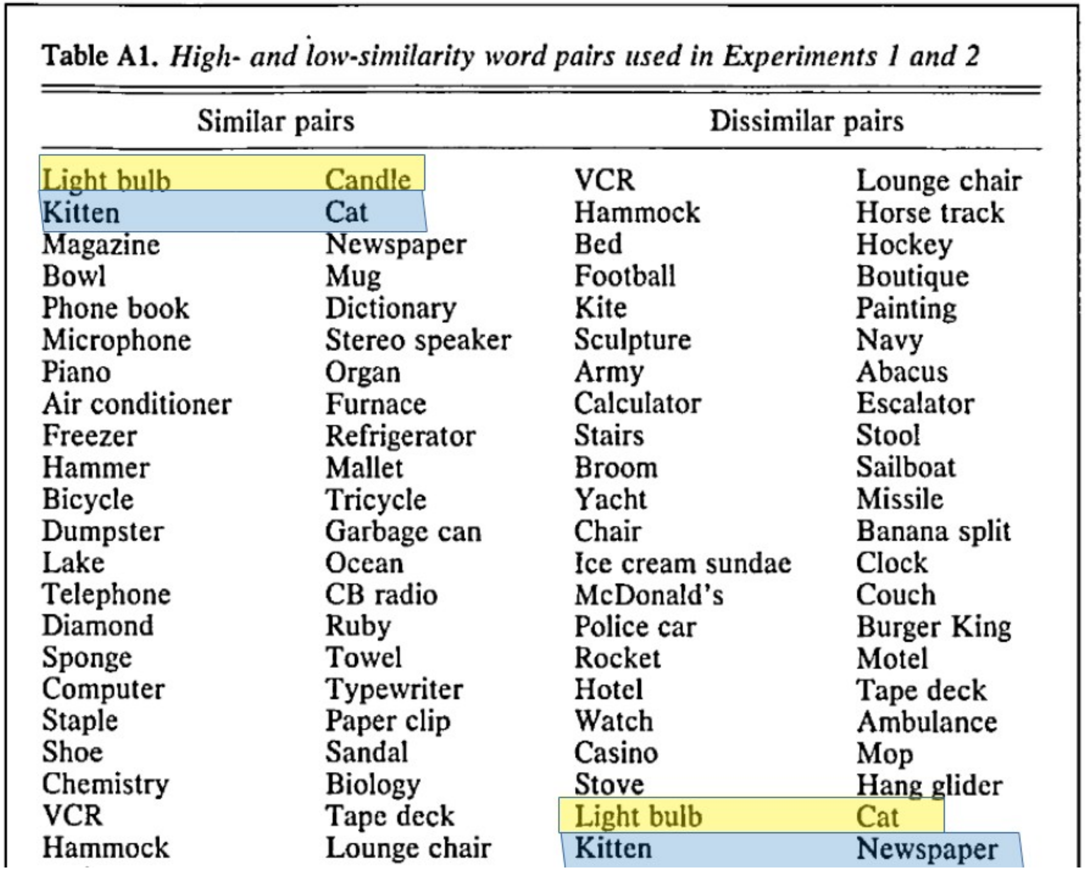


Table A1. *High- and low-similarity word pairs used in Experiments 1 and 2*

Similar pairs		Dissimilar pairs	
Light bulb	Candle	VCR	Lounge chair
Kitten	Cat	Hammock	Horse track
Magazine	Newspaper	Bed	Hockey
Bowl	Mug	Football	Boutique
Phone book	Dictionary	Kite	Painting
Microphone	Stereo speaker	Sculpture	Navy
Piano	Organ	Army	Abacus
Air conditioner	Furnace	Calculator	Escalator
Freezer	Refrigerator	Stairs	Stool
Hammer	Mallet	Broom	Sailboat
Bicycle	Tricycle	Yacht	Missile
Dumpster	Garbage can	Chair	Banana split
Lake	Ocean	Ice cream sundae	Clock
Telephone	CB radio	McDonald's	Couch
Diamond	Ruby	Police car	Burger King
Sponge	Towel	Rocket	Motel
Computer	Typewriter	Hotel	Tape deck
Staple	Paper clip	Watch	Ambulance
Shoe	Sandal	Casino	Mop
Chemistry	Biology	Stove	Hang glider
VCR	Tape deck	Light bulb	Cat
Hammock	Lounge chair	Kitten	Newspaper

Gentner & Markman. Structural alignment in comparison: No difference without similarity. Psychological Science, 5(3):152–158, 1994.

Contrastive Explanations and Causality

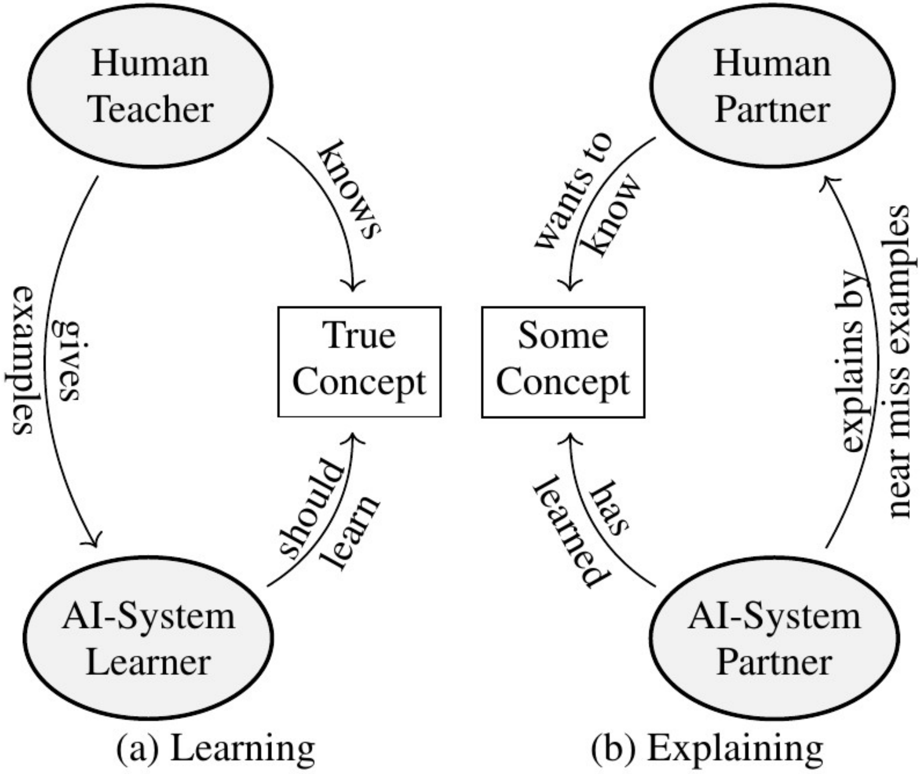
Causal explanations are contrastive (Tim Miller, 2019, refering to P. Lipton, Contrastive explanation, Royal Institute of Philosophy Supplement 27, 1990)

- To be a beetle, an arthropod must have six legs, but this does not cause an arthropod to be a beetle – other causes are necessary.
- But, to answer the question: “Why is image J labelled as a Beetle instead of a Spider?” it is sufficient to cite the fact that the arthropod in the image has six legs.
- We do not need information about eyes, wings, or stingers to answer this, whereas to explain why image J is a spider in a non-contrastive way, we must cite all causes.

Type	No. Legs	Stinger	No. Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	2	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	5	✔	2

Tim Miller, Explanation in artificial intelligence: Insights from the social sciences. AIJ 2019

Near Miss Explanations for Effective Learning and Effective Teaching



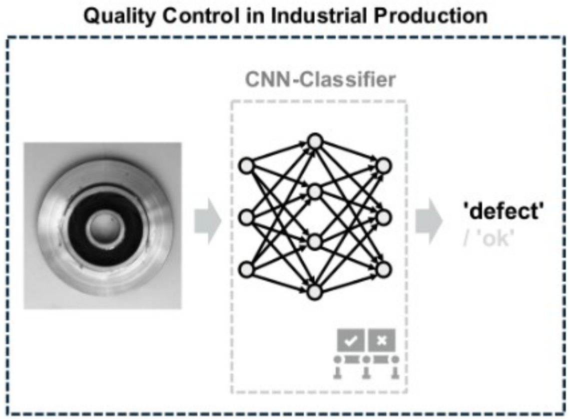
Patrick Winston, Learning structural descriptions from examples.
MIT/LCS/TR-76, 1970.

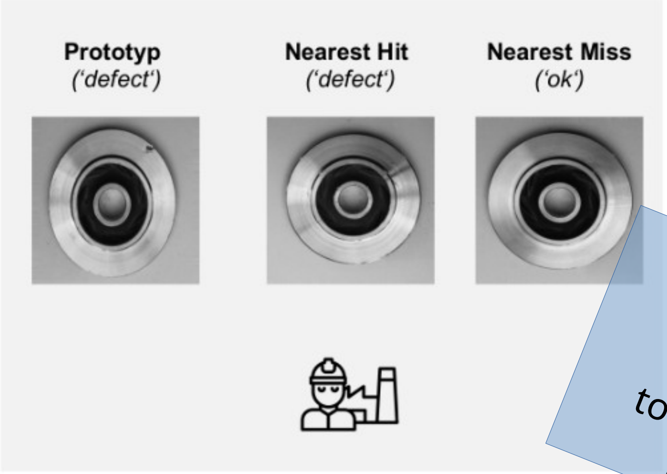
Principles of efficient teaching

Shafto, Goodman, & Griffiths, A rational account of pedagogical reasoning: Teaching by, and learning from, examples. Cognitive Psychology, 71, 55-89, 2014

Telle, J. A., Hernández-Orallo, J., & Ferri, C. (2019). The teaching size: computable teachers and learners for universal languages. Machine Learning, 108(8), 1653-1675.

Example-based Explainable AI (XAI) Demonstrator





Help the quality engineer to understand classification boundaries of the model to provide helpful examples for model adaptation

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

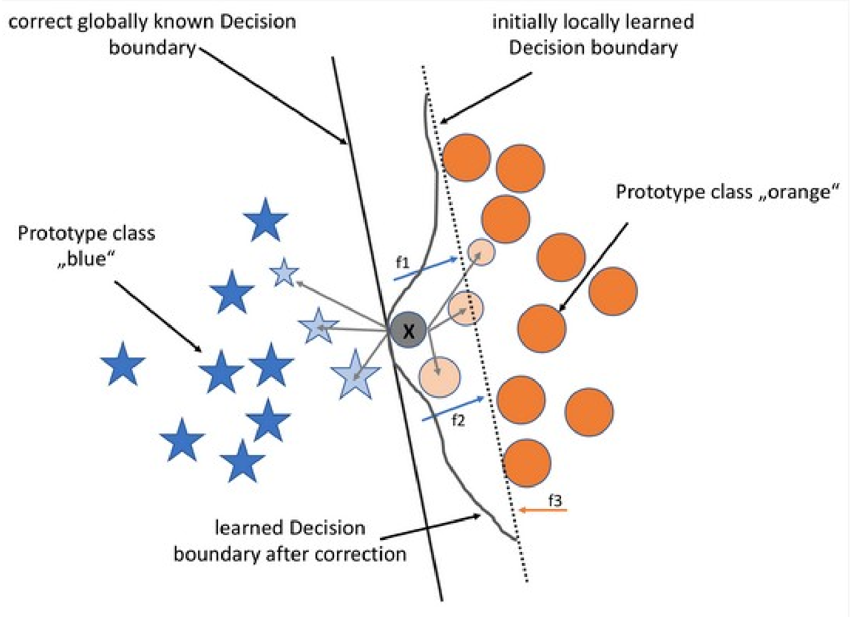
$$\mathrm{MMD}^2(X, Y) := \frac{1}{|X|^2} \sum_{x_1, x_2 \in X} k(x_1, x_2) + \frac{1}{|Y|^2} \sum_{y_1, y_2 \in Y} k(y_1, y_2) - \frac{2}{|X| \cdot |Y|} \sum_{x \in X, y \in Y} k(x, y)$$

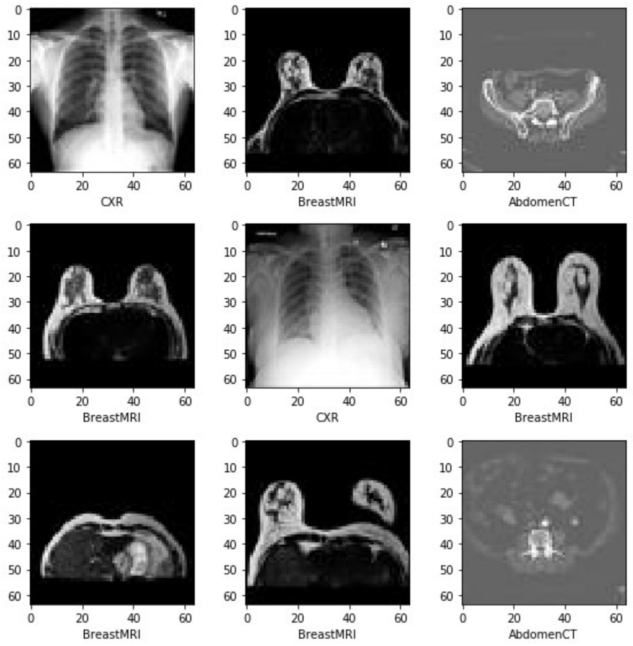
Maximum Mean Discrepancy, similarity measure on distributions

Extended to Near Miss Explanations

Herchenbach, Müller, Scheele, & Schmid, Explaining image classifications with near misses, near hits and prototypes. ICPRAI 2022.

XAI: Explaining by Near-miss Examples





https://www.kaggle.com/datasets/andrewmvd/medical-mnist

Kiefer, Sebastian, Mareike Hoffmann, and Ute Schmid. "Semantic Interactive Learning for Text Classification: A Constructive Approach for Contextual Interactions." Machine Learning and Knowledge Extraction 4.4 (2022): 994-1010.

Slany, Emanuel, et al. "CAIPI in practice: Towards explainable interactive medical image classification." Artificial Intelligence Applications and Innovations. AIAI 2022 IFIP WG 12.5 International Workshops: MHDW 2022, 5G-PINE 2022, AIBMG 2022, ML@ HC 2022, and AIBEI 2022, Hersonissos, Crete, Greece, June 17–20, 2022, Proceedings. Cham: Springer International Publishing, 2022.

Hybrid AI/Neuro-Symbolic AI

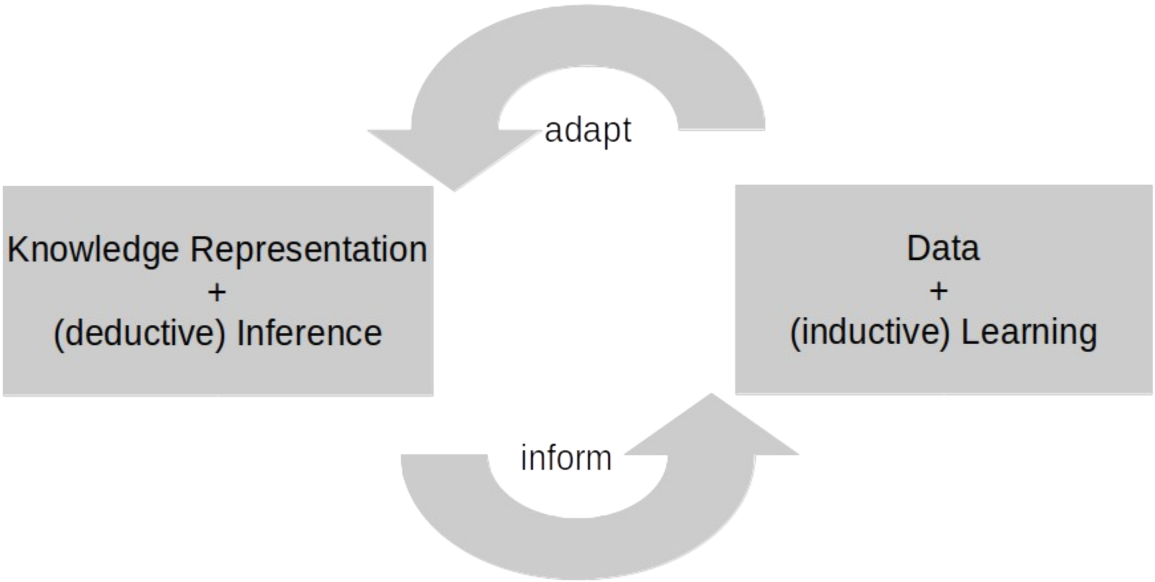


Table 1.
Results for ensemble embeddings with set IoU (sIoU), mean cosine distance to the runs (Cos.d.), and index of conv layer or block (L) (cf. Fig. 3).

		L	sIoU	Cos.d.
AlexNet	NOSE	2	0.228	0.040
	MOUTH	2	0.239	0.040
	EYES	2	0.272	0.058
VGG16	NOSE	7	0.332	0.104
	MOUTH	6	0.296	0.154
	EYES	6	0.350	0.197
ResNeXt	NOSE	6	0.264	0.017
	MOUTH	5	0.237	0.020
	EYES	7	0.302	0.020

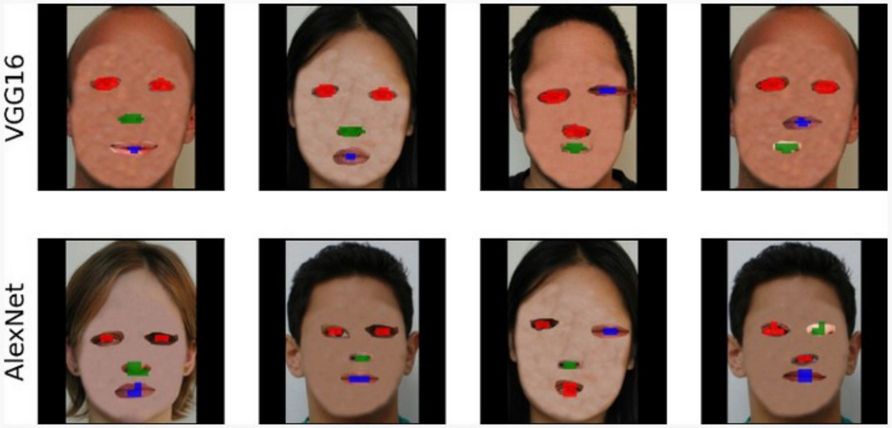


Fig. 4.
Ensemble embedding outputs of NOSE (green), MOUTH (blue), EYES (red). (Color figure online)

Rabold, Schwalbe, Schmid, Expressive Explanations of DNNs by Combining Concept Analysis with ILP, KI 2020

NeurIPS Proceedings

DeepProbLog: Neural Probabilistic Logic Programming

Part of Advances in Neural Information Processing Systems 31 (NeurIPS 2018)

Bibtex Metadata Paper Reviews Supplemental

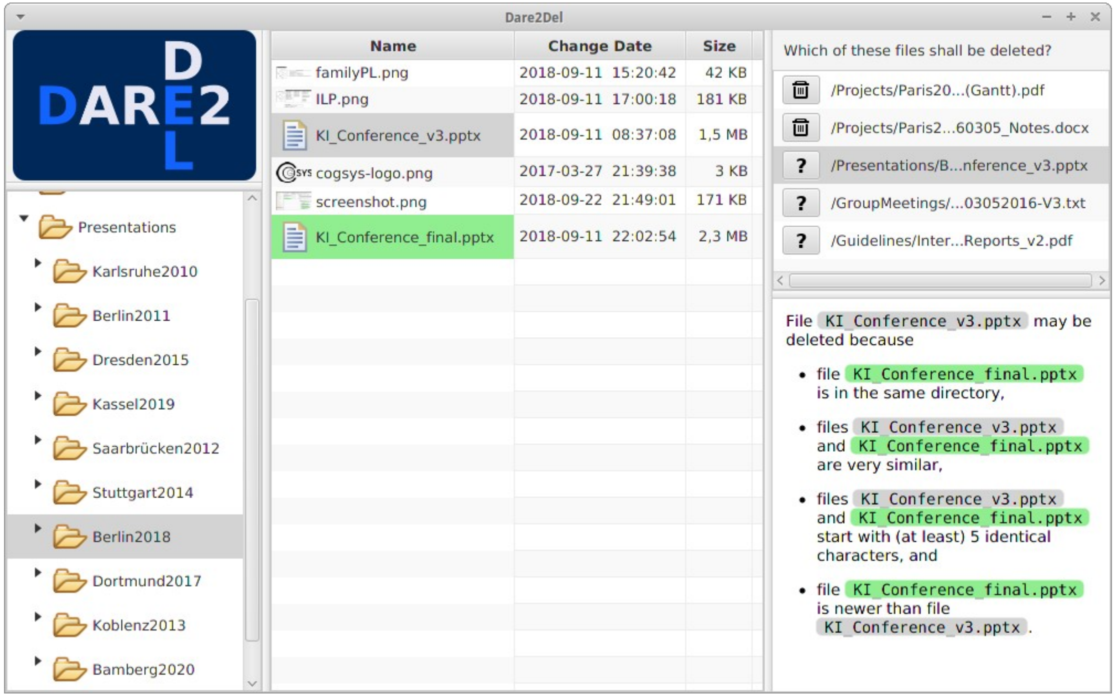
Authors

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, Luc De Raedt

Table 2.
Learned rules for different architectures and their fidelity scores (accuracy and F1 score wrt. to the original model predictions). Learned rules are of common form `face(F):- contains(F, A), isa(A, nose), contains(F, B), isa(B, mouth), distinctPart`

Arch.	Accuracy	F1	Distinct rule part
VGG16	99.60%	99.60%	`top_of(A, B), contains(F, C), top_of(C, A)`
AlexNet	99.05%	99.04%	`contains(F, C), left_of(C, A), top_of(C, B), top_of(C, A)`
ResNext	99.75%	99.75%	`top_of(A, B), contains(F, C), top_of(C, A)`

Deleting Irrelevant Files/Data



Deutsche Forschungsgemeinschaft

What must be minimally changed that this file is not classified as Irrelevant?

Schmid, U. (2021). Interactive learning with mutual explanations in relational domains. In: S. Muggleton and N. Chater, Human-like Machine Intelligence,(chap.~17). 338-354, OUP.

Ultra-Strong Machine Learning

Donald Michie (1988):

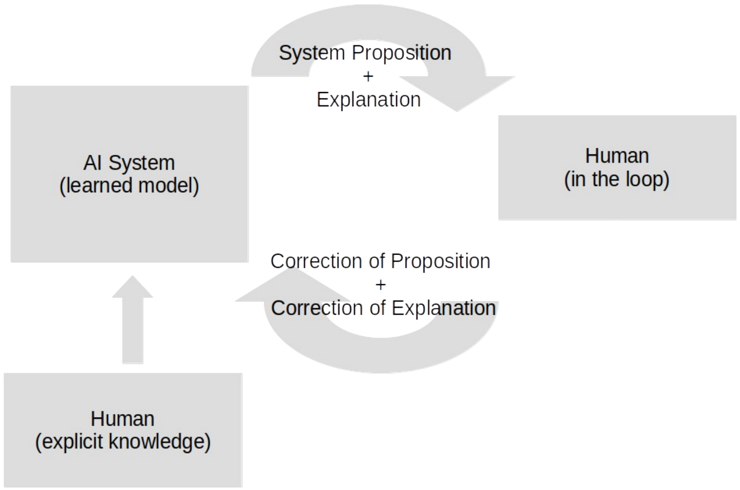
- **Weak ML:** machine learner produces improved predictive performance with increasing amounts of data
- **Strong ML:** additionally requires the learning system to provide its hypotheses in symbolic form (interpretable machine learning, e.g. Rudin, Nature ML, 2019)
- **Ultra-strong ML:** extends the strong criterion by requiring the learner to teach the hypothesis to a human, whose performance is consequently increased to a level beyond that of the human studying the training data alone

Human to Machine Teaching

Machine to Human Teaching

Mutual Human-Machine Explanations

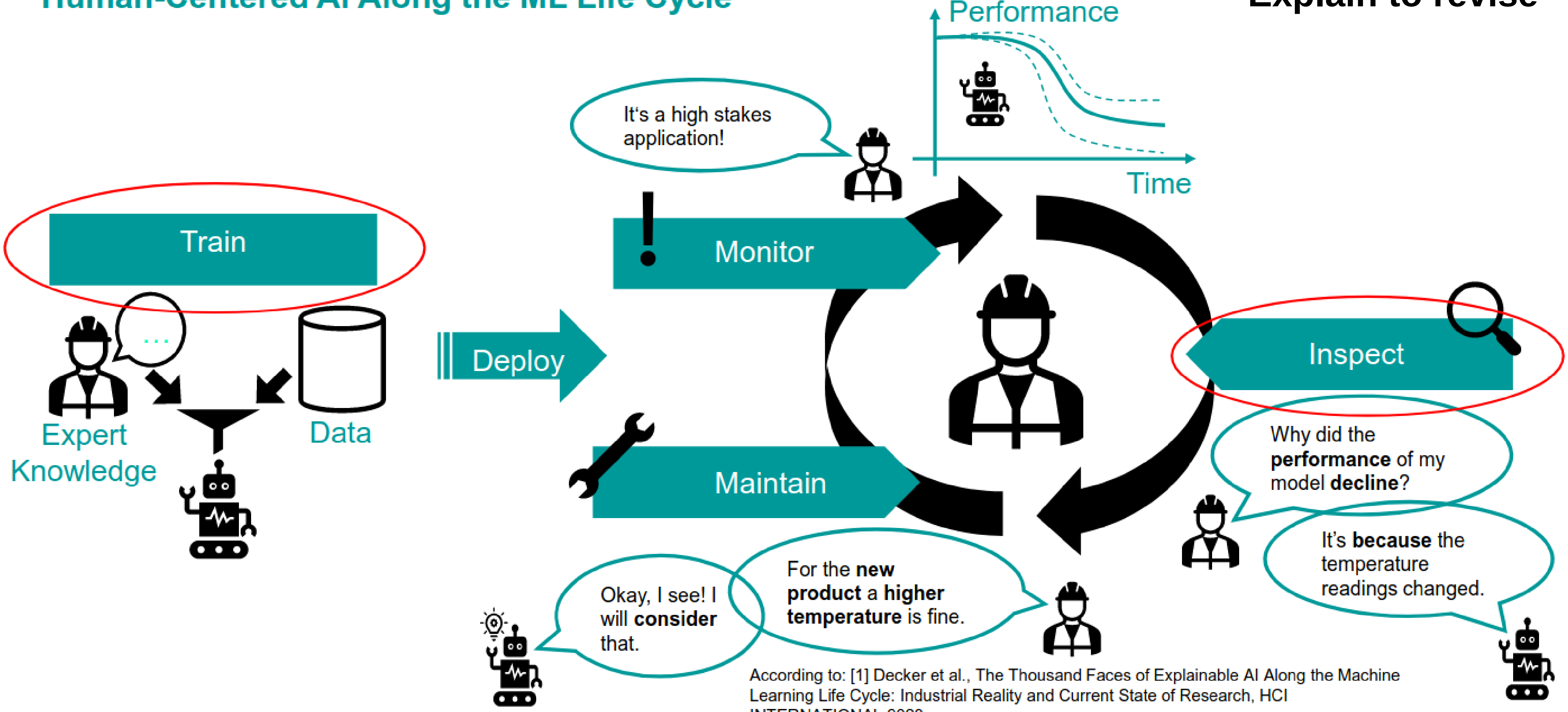
- Human explanation: label correction plus correcting the explanation → model adaptation (*explanatory interactive ML*)
- Advantages of human in the loop:
  - human guidance for ML (expert knowledge, common sense)
  - might also be a cure against automation bias
  - no marginalizaiton of human competences by autnomous AI



- Accountability problem: who is allowed to correct a model decision leading to changes of the system behaviour?

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

Image-based Quality Control of Welding Seams

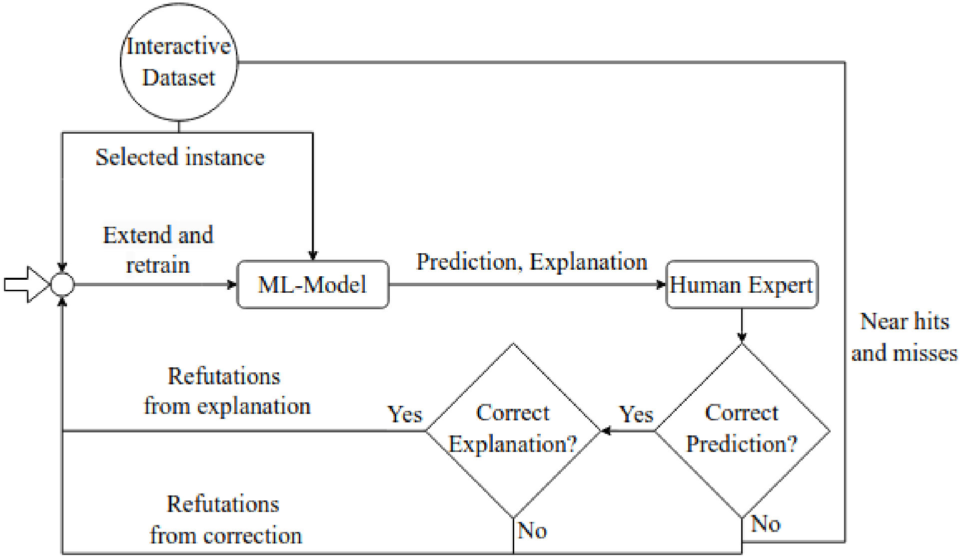
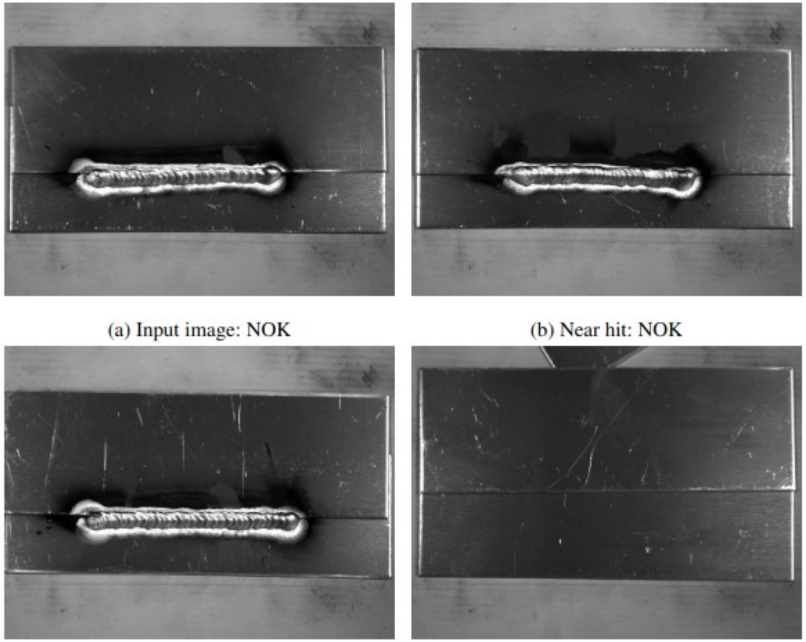


Fig. 3: Illustration of how the human interaction pipeline works. First, an image with the highest potential for information gain is selected. For this image, the AI predicts the class and explains its decision to the human expert. We generate refutations depending on the expert's feedback and add the image and the refutations to the training dataset. If the prediction is wrong, we also expect feedback from the user regarding the nearest hit-and-miss of the image.



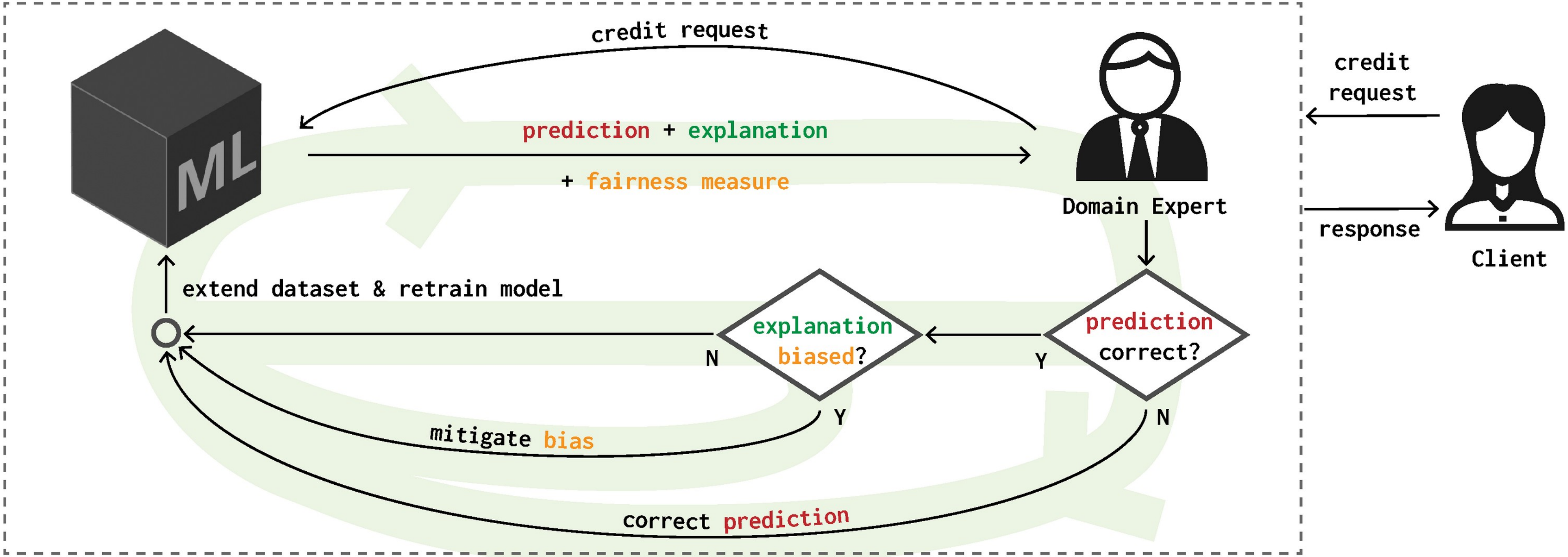
(a) Input image: NOK (b) Near hit: NOK (c) Near miss: OK (d) Furthest hit: NOK

Fig. 4: An example of near hits and misses. (a) First, an image of the input image is presented. It consists of an irregular fish scale, and is therefore labeled as NOK. (b) We select the nearest image with the same label NOK; which is a welding seam that consists of an irregular welding seam and a possible binding error. (c) Additionally we show the nearest image with the label OK. (d) Lastly, we show the image, which is the furthest from our input image, which is a plate with no welding seam present.

Collaboration with Porsche digital,
Gramelt, Höfer, Schmid, Interactive Explainable Anomaly Detection For Industrial Settings, ECCV 2024

FairCAIPI

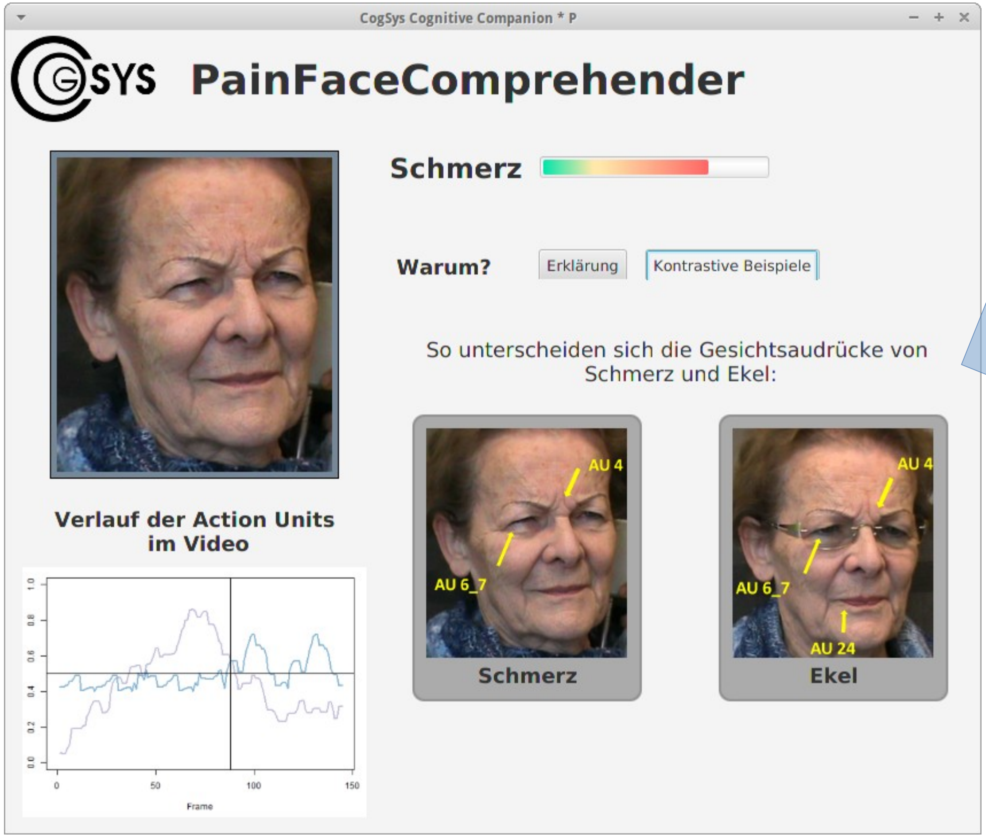
Interactive ML



Heidrich L, Slany E, Scheele S, Schmid U. FairCaipi: A Combination of Explanatory Interactive and Fair Machine Learning for Human and Machine Bias Reduction. *Machine Learning and Knowledge Extraction*. 2023; 5(4):1519-1538.

XAI for Educating Nurses

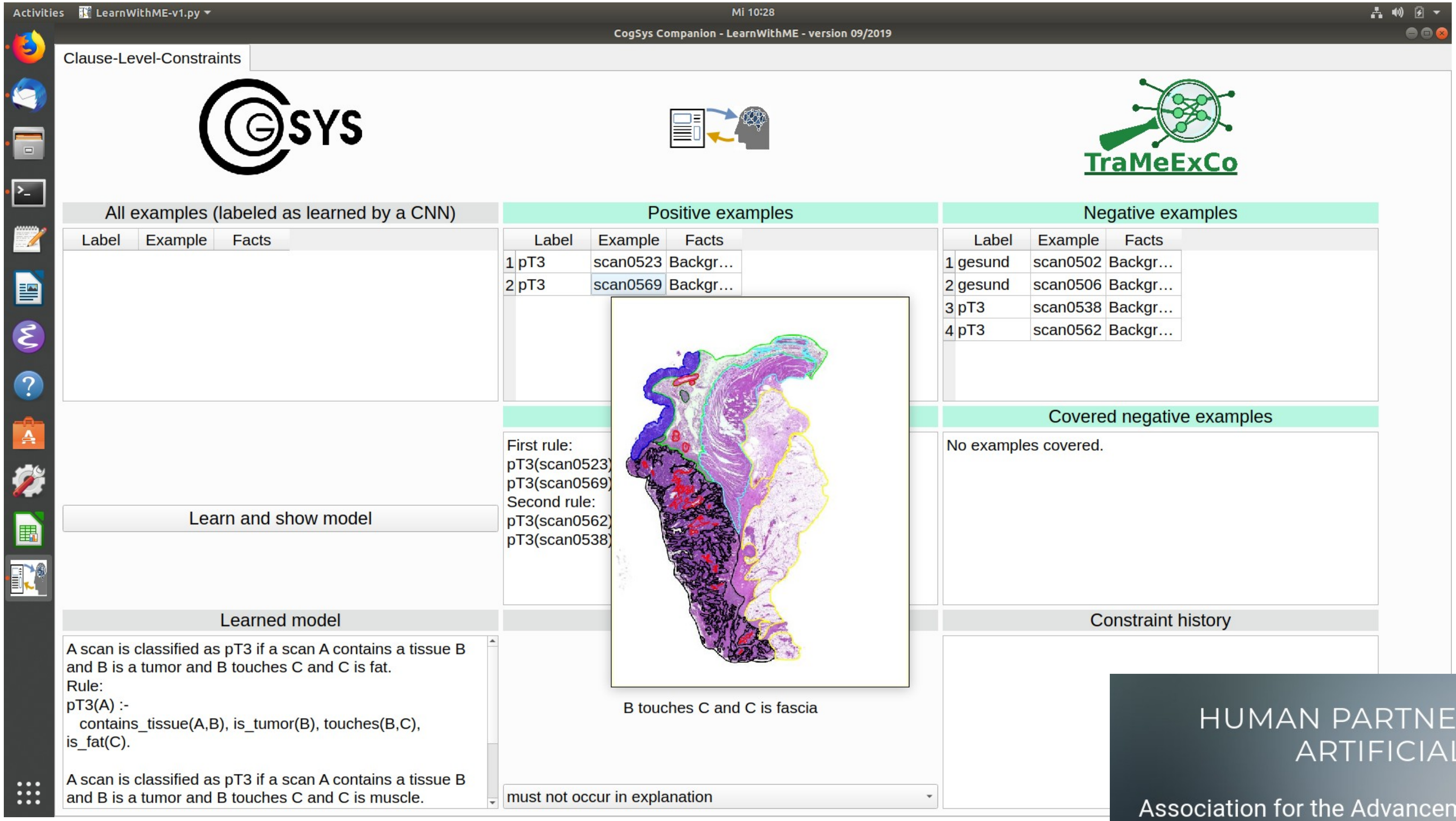
DFG Deutsche Forschungsgemeinschaft



The facial expression of Ms Miller Indicates that she is in pain and not that she is disgusted

Hassan, T., Seuß, D., Wollenberg, J., Weitz, K., Kunz, M., Lautenbacher, S., ... & Schmid, U. (2019). Automatic detection of pain from facial expressions: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(6), 1815-1831.

Knowledge-informed, explainable and interactive ML for Medical Diagnosis



GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Schmid, Ute, and Bettina Finzel. "Mutual explanations for cooperative decision making in medicine." KI-Künstliche Intelligenz 34.2 (2020): 227-233.

Explanation Dialogs

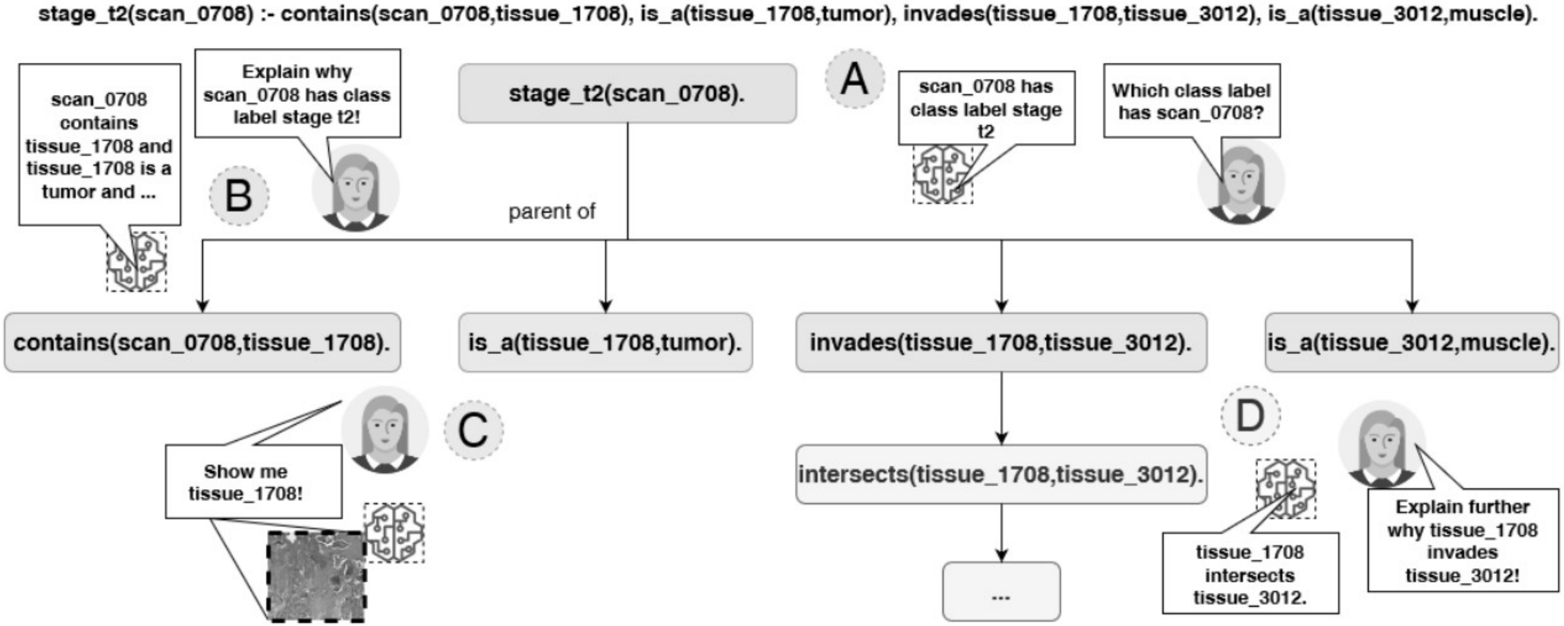


Figure 2: An explanatory tree for *stage_t2(scan_0708)*, that can be queried by the user to get a local explanation why scan_0708 is labeled as T2 (steps A and B). A dialogue is realized by further requests, either to get more visual explanations in terms of prototypes (step C) or to get more verbal explanations in a drill-down manner (step D).

Finzel, Bettina, et al. "Explanation as a process: user-centric construction of multi-level and multi-modal explanations." KI 2021: Advances in Artificial Intelligence: 44th German Conference on AI, Virtual Event, September 27–October 1, 2021, Proceedings 44. Springer International Publishing, 2021.

Take Away

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

- The advance in AI has huge potential for many application domains, among them medical diagnosis, drug design, intelligent production, education
- For trustworthy AI applications, transparency, fairness, and human agency and oversight are crucial
- New challenges for AI research: explainability, knowledge-informed machine learning, fairAI methods, explain to revise methods of interactive machine learning
- The AI Act of the European Union addresses requirements for trustworthy AI, however it has to be seen how these are controlled and enforced (without hindering research and novel applications)

What is your assessment of trustworthiness of current/future AI systems? (after the lecture)

- In what application domains would you trust
  - autonomous AI systems?
  - human-supervised AI systems?
- In what application domains would you not trust AI systems?
- Are there specific AI approaches in which you would put more/less trust?