Reconciling Knowledge-Based and Data-Driven AI for Human-in-the-loop Machine Learning

Ute Schmid

Cognitive Systems Group
University of Bamberg
www.uni-bamberg.de/cogsys

10th Progic, online, Sept. 1-3, 2021

Historical Development of AI Research

- 1st Wave of AI: Exclusive focus on explicit representation of knowledge
- Advantage: Powerful algorithms with provable characteristics
- But: A large amount of human knowledge is not available to inspection and verbalisation (*Polyani's Paradox*)
  - Implicit/tacit knowledge
    e.g., perceptual knowledge, such as object recognition / face recognition
  - Highly automated expert knowledge ("gut feeling")
  - Procedural knowledge / skills
    e.g., driving a bicyle, policy in game playing
  - Common sense reasoning
    e.g., what does not change when performing an action (frame problem)
- 2nd Wave of AI: Exclusive focus on data-intensive machine learning
  - But: high demands on amount and quality of data ("garbage in garbage out")
  - Labeling of training data in specialized domains demands high expertise (medical diagnostics, quality control)

$\hookrightarrow$ From Knowledge Engineering Bottleneck to
Data Engineering Bottleneck

Data Engineering Bottleneck – the next AI winter?

Nuremberg Funnel, 1910; `https://de.wikipedia.org/`

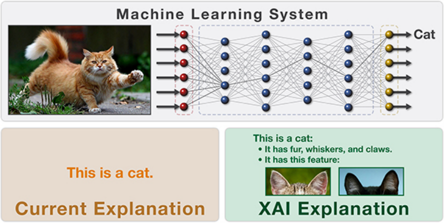
Polanyi's Revenge

(Subbarao Kambhampati, Communications of the ACM, February 2021)

- In AI resesarch as well as practice:
  Polanyi's paradox $\hookrightarrow$ Polanyi's revenge
- Recent advances have made AI synonymous with learning from massive amounts of data, even in tasks for which we do have explicit theories and hard-won causal knowledge!
- Knowledge is injected in deep learning through architectural biases and carefully manufactured examples
- Anecdotal evidence: industry practitioners readily convert doctrine and standard operating procedures into 'data' only to have the knowledge be 'learned back' from that data.

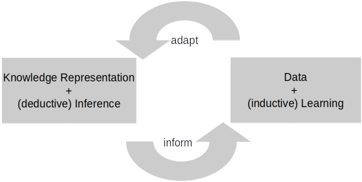
Figure. "Human, grant me the serenity to accept the things I cannot learn, learn the things I can, and wisdom to know the difference."

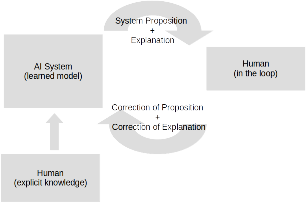
3rd Wave of AI: Explainable AI (XAI)
Hybrid, explanatory, interactive, human-centric



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

David Gunning, IJCAI 2016

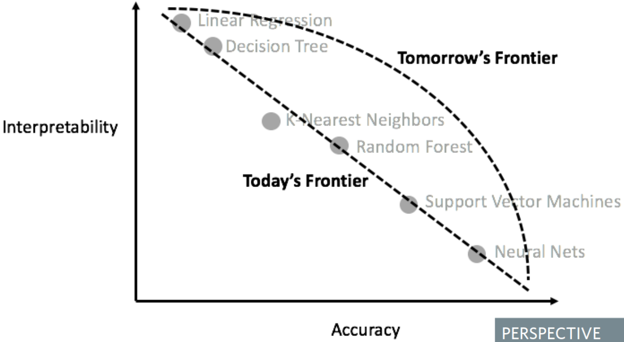




Outline

- On to the 3rd Wave of AI
  - 1st Wave: knowledge-based
  - 2nd Wave: data-driven
  - 3rd Wave: hybrid, XAI, human-centric
- Inductive (Logic) Programming
  - Natural Combination of Learning and Reasoning in First Order Logic
  - Learning in Relational Domains
  - Expressive Approach to Intrinsically Interpretable Machine Learning
  - Neural-symbolic Integration (CNN + ILP)
- Explanatory and Interactive Machine Learning
  - The Need for Multi-Modal Explantions
  - Empirical Evidence for Effects of Explanations on Performance and Trust
  - Mutual Explanations in Human-AI Partnerships (Domain Experts)
  - Explanations for Novices – Intelligent Tutor Systems

Predictive Accuracy & Comprehensibility of Models/Decisions



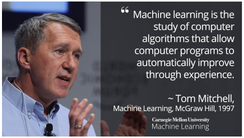
PERSPECTIVE
https://doi.org/10.1038/s42256-019-0048-x
nature machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Machine Learning – A Research Area with Long Tradition

- At the beginning (in accordance with goals of early AI): human-like machine learning – *A computer algorithm analyses data and creates a general rule it can follow and discard unimportant data.*
  - Arthur Samuel (1952) – learning a strategy for checkers
  - Donald Michie (1963) – reinforcement learning for Tic-tac-toe
  - Tom Mitchell (1977) – version spaces
  - Patrick Winston (1981) – relational learning with near misses
  - Gerald de Jong (1982) – explanation-based generalization
  - Ryszard Michalski (1983) – concept learning
  - Ross Quinlan (1986) – decision trees
  - Pat Langley (1988) – learning from problem solving experience
  - Stephen Muggleton (1991) – inductive logic programming

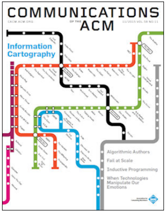


www.ibm.com/ibm/history/ibm100/images/icp/, miro.medium.com

ILP: Learning Prolog Programs

- Hypotheses/models are represented as Prolog programs
- Examples are presented by target predicates (positive and negative) and by background knowledge
- In some approaches: also by background theories
- $\hookrightarrow$ Uniform representation as Horn clauses

Gulwani, Hernandez-Orallo, Kitzelmann, Muggleton, Schmid, Zorn, Inductive Programming meets the real world, CACM 58(11), 2015



Machine Learning

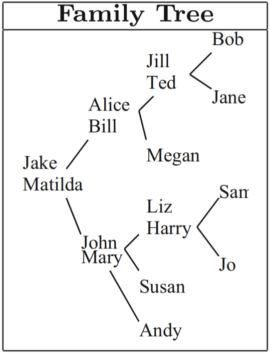
July 2018, Volume 107, Issue 7, pp 1119–1140 | Cite as

Ultra-Strong Machine Learning: comprehensibility of programs learned with ILP

Authors　Authors and affiliations

Stephen H. Muggleton, Ute Schmid, Christina Zeller, Alireza Tamaddoni-Nezhad, Tarek Besold

Example: Family Domain




```
% Background Knowledge
father(jake,bill).    mother(matilda,bill).
father(jake,john).    mother(matilda,john).
father(bill,ted).     mother(alice,jill).
father(bill,megan).   mother(alice,ted).
father(john,harry).   mother(alice,megan).
father(john,susan).   mother(mary,harry).
father(ted,bob).      mother(mary,susan).
father(ted,jane).     mother(mary,andy).
father(harry,san).    mother(jill,bob).
father(harry,jo).     mother(jill,jane).
mother(liz,san).      mother(liz,jo).
```

```
% Examples
grandparent(matilda,megan).    not grandparent(megan,matilda).
grandparent(matilda,harry).    not grandparent(jake,jake).
grandparent(jake,susan).       not grandparent(matila,alice).

% Learned hypothesis    (parent can be background theory or invented)
grandparent(X,Y) :- parent(X, Z), parent(Z,Y).
parent(X,Y) :- father(X,Y).
parent(X,Y) :- mother(X,Y).
```

```
% Background Theory for Spatial Relations
% --------------------------------------
% Area X touches area Y if holds that they have at least one boundary point
% in common, but no interior points.
touches(X,Y) :- I is intersection(X,Y), not(empty(I)),
InteriorX is interior(X), InteriorY is interior(Y),
J is intersection(InteriorX,InteriorY), empty(J).
% disjoint(X,Y) :- ...
% includes (X,Y) :- ...
% ...
% positive examples for diagnostic class pT3
% ----------------------------------------
% scan123 is classified as pT3. The scan is composed of areas of
% different tissues such as fat and tumor which are in specific spatial relations.
pt3(scan123).
contains_tissue(scan123,t1). contains_tissue(scan123,f1).
contains_tissue(scan123,f2).
is_tumor(t1). is_fat(f1). is_fat(f2)
touches(t1,f1). disjoint(f1,t1).
% negative examples for diagnostic class pT3 (e.g. pT2, pT4)
% ----------------------------------------
% ...
% Induced Rules: (learned from data with ILP)
% ----------------------------------------
% A scan is classified as pT3 if a scan A contains a tissue B
% and B is a tumor and B touches C and C is fat.
pT3(A) :-
   contains_tissue(A,B), is_tumor(B), is_fat(C), touches(B,C).
% further rules ...
```

Bruckert, Finzel, Schmid, The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions, Frontiers in AI, 2020

ILP Algorithms

Given a tuple (B, E^+, E^-) where:

- B denotes background knowledge
- E^+ denotes positive examples of the concept
- E^- denotes negative examples of the concept

An ILP algorithm returns a hypothesis $H \in \mathcal{H}$ such that:

$$\forall e \in E^+, H \cup B \vdash e \text{ (i.e. H is complete)}$$

$$\forall e \in E^-, H \cup B \nvdash e \text{ (i.e. H is consistent)}$$

- FOIL (Quinlan, 1990): Generate-and-test, sequential covering (ID3, C4.5, simulteneous covering by the same author)
- Golem, Progol, Aleph, Metagol (Muggleton, since 1990ies): learning from entailment in different variants
- Igor (Kitzelmann & Schmid, JMLR 2006; Schmid & Kitzelmann, CSR 2011): Inductive (functional) programming
- ProbLog (de Raedt, 2007): combining logical and statistical learning

Algorithm

FOIL(*Target_predicate*, *Predicates*, *Examples*)

- *Pos* ← those *Examples* for which the *Target_predicate* is *True*
- *Neg* ← those *Examples* for which the *Target_predicate* is *False*
- *Learned_rules* ← {}
- while *Pos*, Do
  - *NewRule* ← the rule that predicts *Target_predicate* with no precondition
  - *NewRuleNeg* ← *Neg*
  - while *NewRuleNeg*, Do
    - *Candidate_literals* ← generate new literals for *NewRule*, based on *Predicates*
    - *Best_literal* ← $argmax_{L \in Candidate\_literals} FoilGain(L, NewRule)$
    - add *Best_literal* to preconditions of *NewRule*
    - *NewRuleNeg* ← subset of *NewRuleNeg* that satisfies *NewRule* preconditions
  - *Learned_rules* ← *Learned_rules* + *NewRule*
  - *Pos* ← *Pos* − {members of *Pos* covered by *NewRule*}
- Return *Learned_rules*

Probabilistic Inductive Logic Programming

- Statistical Relational Learning (StarAI)
- Motivation: Biological Graphs
  `path(gene_620, disease_altzheimer)`
  edges are typically probabilistic

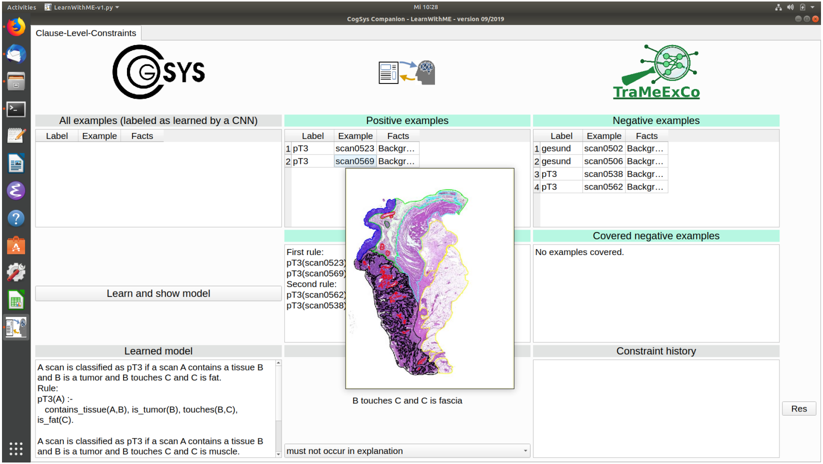
Example 1 *As an example, consider:*

1.0: likes(X,Y):- friendof(X,Y).
0.8: likes(X,Y):- friendof(X,Z), likes(Z,Y).
0.5: friendof(john,mary).
0.5: friendof(mary,pedro).
0.5: friendof(mary,tom).
0.5: friendof(pedro,tom).

De Raedt, Kimmig, Toivonen, ProbLog: A probabilistic Prolog and its application in link discovery, IJCAI 2007

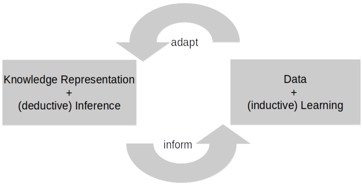
XI-ML for Medical Diagnosis

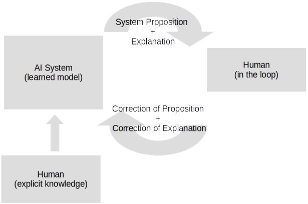
GEFÖRDERT VOM Bundesministerium für Bildung und Forschung



Explicit and Implicit Knowledge Injection in ML

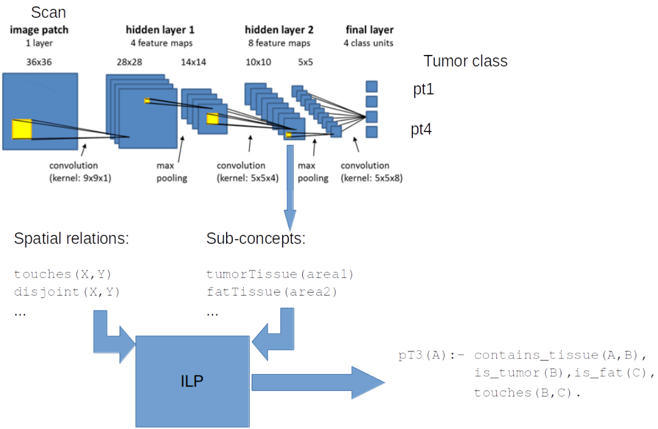
- Combine with KR whenever explicit knowledge is available, e.g., domain specific/expert knowledge
- Take into account formal approaches for common sense/world knowledge, e.g., temporal or spatial calculi
- Human experts might not be able to explicitly formulate all rules necessary to perform a diagnosis – but, they recognize errors and can correct them
  $\hookrightarrow$ interactive learning





Neural-symbolic Integration

- Many recent approaches (de Raedt et al., IJCAI 2020 Survey)
- Combining learning for perceptual domains and interpretable ML
- Blackbox classifiers as sensors



Picasso Faces

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).

Arch.		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

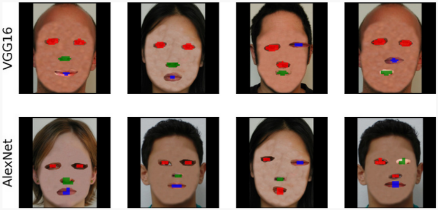


Fig. 4.
Ensemble embedding outputs of NOSE (green), MOUTH (blue), EYES (red). (Color figure online)

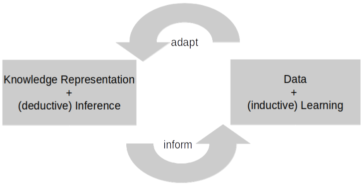
Table 2.
Learned rules for different architectures and their fidelity scores (accuracy and F1 score wrt. to the original model predictions). Learned rules are of common form `face(F):- contains(F, A), isa(A, nose), contains(F, B), isa(B, mouth), distinctPart`

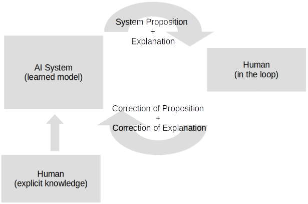
Arch.	Accuracy	F1	Distinct rule part
VGG16	99.60%	99.60%	`top_of(A, B), contains(F, C), top_of(C, A)`
AlexNet	99.05%	99.04%	`contains(F, C), left_of(C, A), top_of(C, B), top_of(C, A)`
ResNext	99.75%	99.75%	`top_of(A, B), contains(F, C), top_of(C, A)`

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

Explicit and Implicit Knowledge Injection in ML

- Combine ML with KR whenever explicit knowledge is available, e.g., domain specific/expert knowledge
- Take into account formal approaches for common sense/world knowledge, e.g., temporal or spatial calculi
- Human experts might not be able to explicitly formulate all rules necessary to perform a diagnosis – but, they recognize errors and can correct them
  $\hookrightarrow$ interactive learning

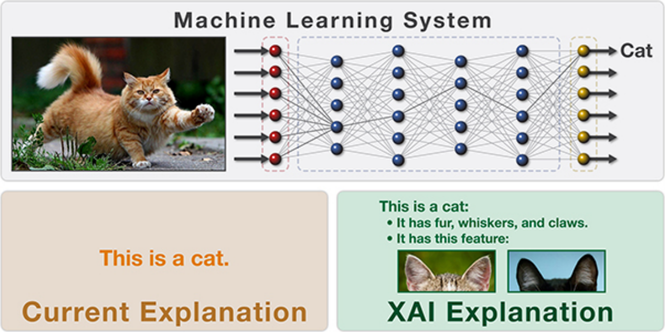




Interactive ML and Explanations

- Interactive ML allows to make use of knowledge in learning
  - More knowledge means less data are necessary
    We do not need to learn what we already know
  - Knowledge can constrain and guide model induction
  - When ground truth labeling is expensive or not available, label corrections might be helpful
- For effective knowledge injection, humans must comprehend (aspects of) the learned model
  - Local explanations to make decisions for specific instances comprehensible
  - Global explanations to make the model itself transparent
- A model might be right for the wrong reason (e.g. Teso & Kersting, 2019)
  $\hookrightarrow$ extend interactive learning to correction of explanations (e.g. Schmid & Finzel, Mutual explanations, KI 2020)

Explainable Artificial Intelligence (XAI)

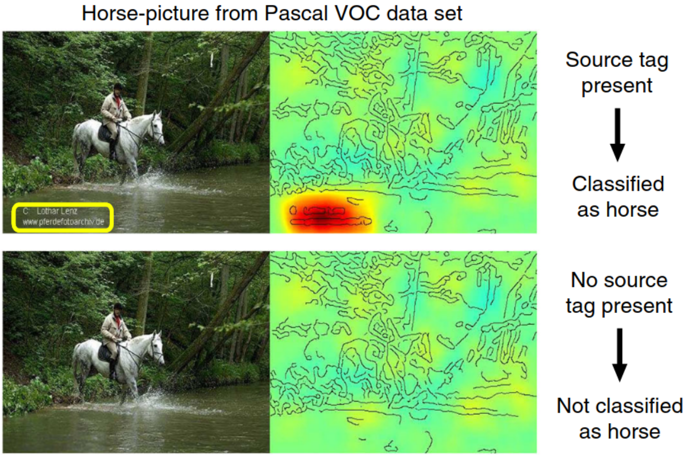


http://www.darpa.mil/program/explainable-artificial-intelligence

David Gunning, IJCAI 2016

First years – nearly exclusive focus on visual explanations (saliency maps): LIME, LRP, Grad-CAM

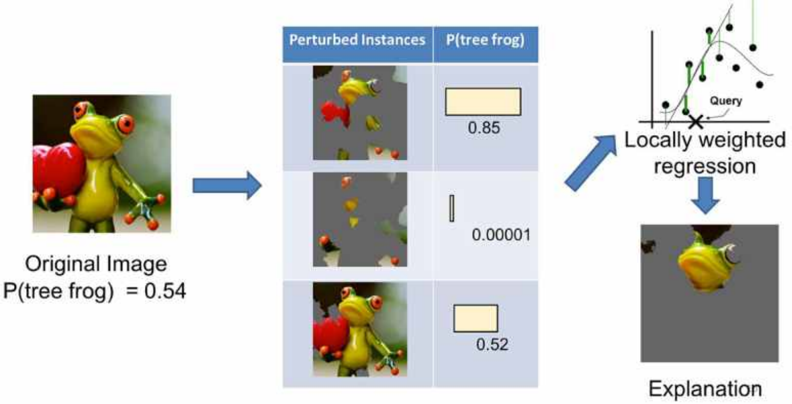
Unmasking Clever Hans Predictors



(Lapuschkin et al., 2019, LRP)

LIME

"Perturbed" samples (deleting part of information, e.g., superpixels, words)

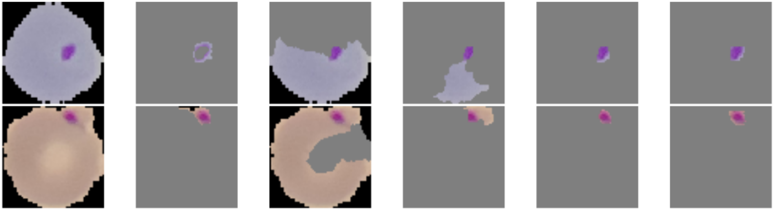


Ribeiro, Singh, Guestin, Why Should I Trust You?: Explaining the Predictions of Any Classifier, KDD 2016

LIME's Superpixel Approach Quick-Shift

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**

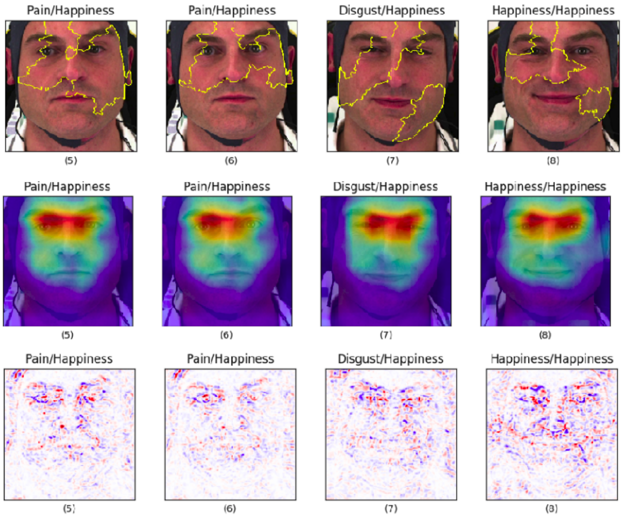


(a) Original (b) Felzenszwalb (c) Quick-Shift (d) Quick-Shift opt. (e) SLIC (f) Compact-Watershed

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

Schallner, Rabold, Scholz, Schmid, Effect of Superpixel Aggregation on Explanations in LIME – A Case Study with Biological Data, AIMLA 2019

Visual Explanations



LIME

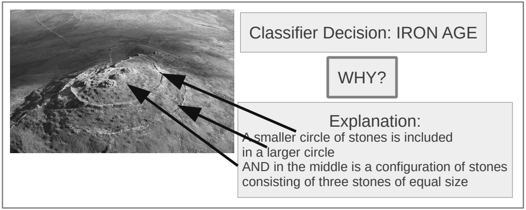
CAM

LRP

Weitz, Hassan, Schmid, Garbas, Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods, tm-Technisches Messen, 2019

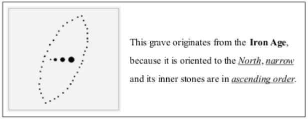
Visual explanations are often not sufficient

- Helpful to recognize overfitting
- Fast communication of information (attention, relevance)
- BUT – visual highlighting is not expressive enough for
  - spatial relations (the blowhole is **on** a supporting part)
  - quantification (**all** blowholes are smaller than 1 mm)
  - feature values (the eyes are **shut** not open)
  - negation (there is **not** a blowhole but a hairline crack)
  - recursion (an arbitraty number of objects of increasing size)



Rabold, Siebers, Schmid, ILP 2018; Rabold, Deininger, Siebers, Schmid, Enriching Visual with Verbal Explanations for Relational Concepts – Combining LIME with Aleph, AIMLA 2019

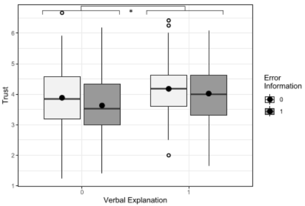
Experimental Findings on Explanations, Joint Performance, and Trust

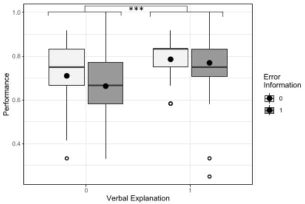


(1) **North** ∧ **Narrow** → ***Iron***
(2) **North** ∧ **Ascending** → ***Iron***
(3) **Narrow** ∧ **Ascending** → ***Iron***

(4 features, based on Medin & Schaffer, 1978, stimuli pattern for classification learning)

		Verbal explanations	
		available	not available
System error information	available	*EG 1*	*EG3*
	not available	*EG 2*	*CG*





(ANOVA, n = 190; a priori power analysis for a medium effect size ($f = .25, \alpha = .05, 1 - \beta = .90$) gave a minimum required sample size of 171 participants)

Thaler & Schmid, Explaining Machine Learned Relational Concepts in Visual Domains – Effects of Perceived Accuracy on Joint Performance and Trust, CogSci2021

Some Observations on Explanations

Tim Miller, AIJ 2019; Tania Lombrozo, TiCS 2006

- There are different possibilities to explain something to someone
  - verbal (different degrees of detail)
  - visual (maybe with symbolic annotations)
  - prototypical examples
  - contrastive (near miss) example
- There is no one-size fits all (context specificity)
- Explanations can be wrong (*right for the wrong reasons*, Teso & Kersting, AAAI/ACM Conference on AI, Ethics, and Society, 2019)
- Explanations are not always helpful (*Beneficial and Harmful Explanatory Machine Learning*, Ai, Muggleton, . . ., Schmid, MLJ 2021)
- Explanations might lead to unjustified trust
- Explanations can be mutual and extend interactive learning (Schmid & Finzel, KI 2020)
- Explanations are a process

Contrastive Near-miss Explanations: Structural Alignment

APPENDIX

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, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. Psychological Science, 5(3), 152-158.)

Contrastive Near-miss Explanations: Relational Learning

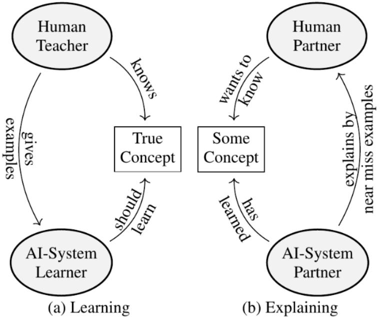


Fig. 2: Duality of learning and explaining

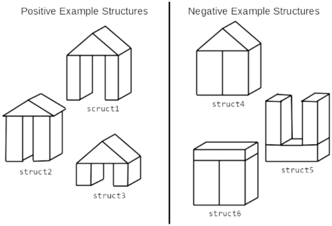


Fig. 5: The positive and negative example structures for the Winston arches domain.

(Rabold, Siebers, Schmid, MLJ, to appear)

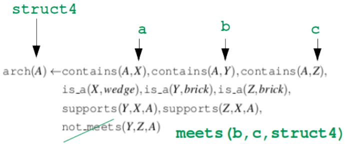
Contrastive Near-miss Explanations: Relational Learning

Local explanation

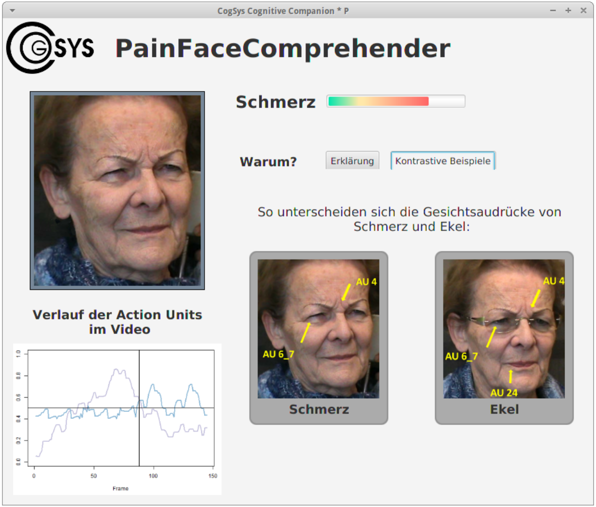
A local explanation for a positive example P is a ground clause $C\theta$ where $C \in T$ such that $P = head(C\theta)$ and $T \models body(C\theta)$.

Near Miss Explanation

Given a local explanation $C\theta$ and a minimally changed clause C' with substitution θ', we call $C'\theta'$ a near miss explanation and $\Delta head(C'\theta')$ a near miss example if $T \models body(C'\theta')$, $T \not\models head(C'\theta')$.



Explanation of Critical Features by Contrastive Alignment



Intelligent Tutor System for Nurses (DFG, PainFaceReader)

Explanatory Dialogue

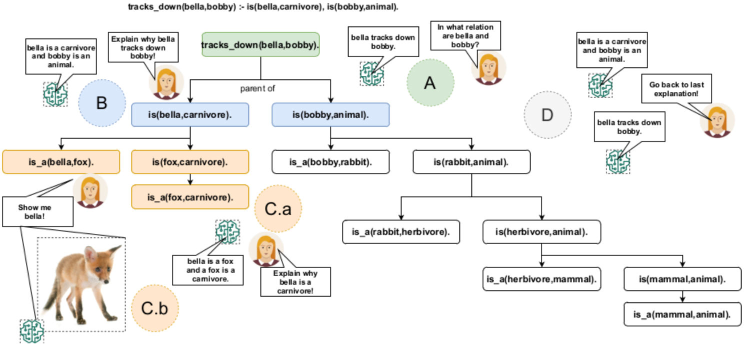


Fig. 8: An explanatory tree for the positive example `tracks_down(bella,bobby)`, that can be queried by the user to get a *local* explanation why Bella tracks down Bobby (reference A and B). A dialogue is realized by different *drill-down* questions, either to get more detailed verbal explanations or visual explanations (references C.a) and C.b)). Furthermore, the user can return to the last explanation (reference D).

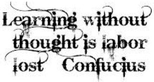
(Finzel, Tafler, Scheele, Schmid, Explanation as a process: user-centric construction of multi-level and multi-modal explanations, KI 2021)

Example Projects

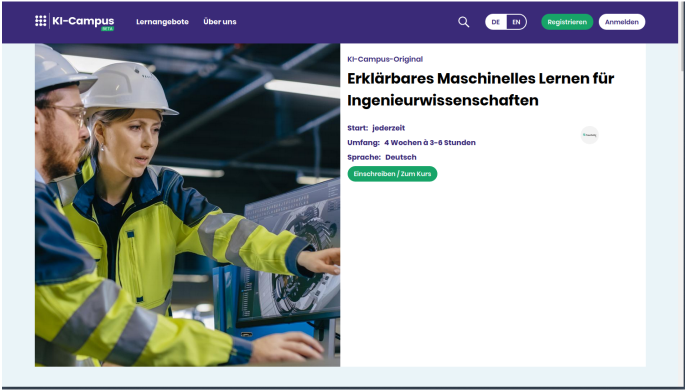
- Transparent Medical Expert Companions (TraMeExCo, BMBF, 2018-2021)
- Video-based automated pain detection exploiting compositional and temporal characteristics of action units (PainFaceReader, DFG, 2018-2021)
- Learning to Delete: Forgetting of Digital Objects as Collaborative Task of Human and AI (Dare2Del, DFG Priority Program Intentional Forgetting (2016-2019, 2019-2022)
- Mensch-KI-Partnerschaft für die proaktive Qualitätskontrolle in der industriellen Fertigung am Beispiel der Wertschöpfungskette der Produktion des elektrischen Antriebsstrangs für die E-Mobilität (KIProQua, BayVFP – Digitalisierung, Start Oct. 2021)
- Human-AI-Partnerships for Explanations in Complex Socio-technical Systems (bidt, with A. Pretschner and E. Hilgendorf, since 2020)

Take Away

- Many application domains have requirements which cannot be met by data intensive blackbox approaches of machine learning alone
- Combining deep learning and ILP supports learning of classifiers for image data together with relational explanations
- Mutual explanations and interactive learning allow to integrate expert/common sense knoweldge into the learning process resulting in less need for data and allowing to correct errouneos decisions of the learned model
- Explanations are not one size fits all – therefore research should address different explanation modalities, their combination, and strategies to select the most helpful explanations
- Research on explanations and human-in-the loop ML requires interdisciplinary collaboration with psychology and education



Learn more about XAI



STIFTERVERBAND DFKI HPI Hasso Plattner Institut mmb Institut NEOCOSMO

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

Thanks to Team, Cooperation Partners, Funding Agencies

Bundesministerium
für Bildung
und Forschung

TraMeExCo (ML-3)

DFG
Deutsche
Forschungsgemeinschaft

Dare2Del
(SPP 1921)

PainFaceReader

Fraunhofer
IIS
Projektgruppe
Comprehensible AI

bidt
Bayerisches
Forschungsinstitut für
Digitale Transformation

Stephen Muggleton (Imperial College)

Cornelia Niessen (FAU)

Stephan Lautenbacher (Uni BA)

Andreas Holzinger (TU Graz)

Torben Weiss (Uni Duisburg-Essen)

Bettina Finzel (XI-ML), Johannes Rabold (Neural-symbolic ML), Mark Gromowski (Logik + Learning), Durgesh Nandini (Interactive ILP), Sarem Seitz (Statistical ML), Gesina Schwalbe (Continental), Deniz Neufeld (Bosch), Sebastian Bruckert (Datev), Dr. Stephan Scheele (embedded AI), Ines Rieger, Dominik Seuß, Jaspar Pahl, Teena Hassan (IIS), Oliver May (IIS Dresden)

FELI
Forschungsgruppe
Elementarinformatik