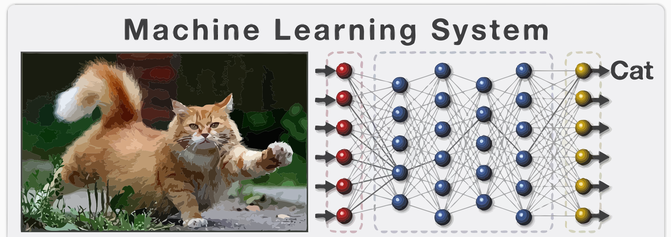
Towards Trustable Explainable AI

Alexey Ignatiev

MONASH University

eXplainable AI



This is a cat.

Current Explanation

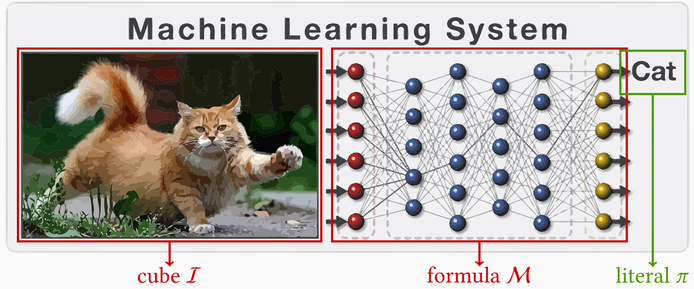
This is a cat:
- It has fur, whiskers, and claws.
- It has this feature:

XAI Explanation

©DARPA

Explanations are important and in some cases **required by law**!

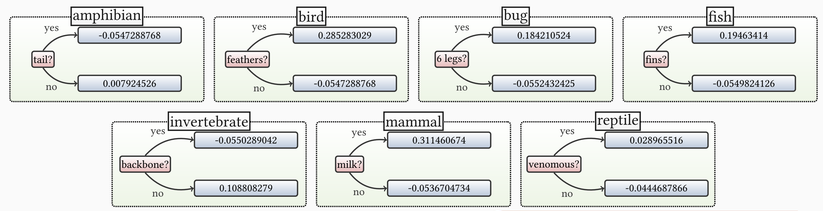
Trustable Explanations



$$\mathcal{I} \land \mathcal{M} \models \pi$$

Abductive explanation $\mathcal{E} \subseteq \mathcal{I}$ is a ***prime implicant*** of $\mathcal{M} \rightarrow \pi$

Heuristic Status Quo



instance:

IF (animal_name = pitviper) ∧ ¬hair ∧ ¬feathers ∧ eggs ∧ ¬milk ∧ ¬airborne ∧ ¬aquatic ∧ breathes ∧ ¬toothed ∧ backbone ∧ predator ∧ venomous ∧ ¬fins ∧ (legs = 0) ∧ tail ∧ ¬domestic ∧ ¬catsize
THEN (class = reptile)

Anchor explains:

IF ¬hair ∧ ¬milk ∧ ¬toothed ∧ ¬fins
THEN (class = reptile)

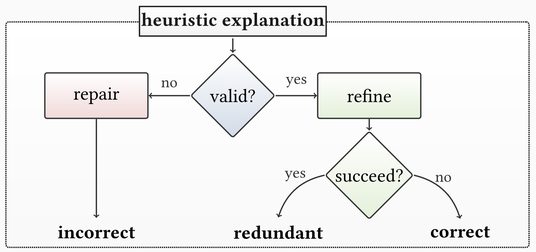
given $\mathcal{E}_h$, check $\mathcal{E}_h \models (\mathcal{M} \rightarrow \pi)$

if **invalid**, $\mathcal{E}_h \land \mathcal{M} \land \neg\pi$ ***is satisfiable***

counterexample to explanation:

IF (animal_name = toad) ∧ ¬hair ∧ ¬feathers ∧ eggs ∧ ¬milk ∧ ¬airborne ∧ ¬aquatic ∧ ¬predator ∧ ¬toothed ∧ backbone ∧ breathes ∧ ¬venomous ∧ ¬fins ∧ (legs = 4) ∧ ¬tail ∧ ¬domestic ∧ ¬catsize
THEN (class = amphibian)

Assessing Explanation Validity



Heuristic Explanations Assessed

Dataset	(# unique)	Explanations: incorrect			redundant			correct		
		LIME	Anchor	SHAP	LIME	Anchor	SHAP	LIME	Anchor	SHAP
adult	(5579)	61.3%	80.5%	70.7%	7.9%	1.6%	10.2%	30.8%	17.9%	19.1%
lending	(4414)	24.0%	3.0%	17.0%	0.4%	0.0%	2.5%	75.6%	97.0%	80.5%
rcdv	(3696)	94.1%	99.4%	85.9%	4.6%	0.4%	7.9%	1.3%	0.2%	6.2%
compas	(778)	71.9%	84.4%	60.4%	20.6%	1.7%	27.8%	7.5%	13.9%	11.8%
german	(1000)	85.3%	99.7%	63.0%	14.6%	0.2%	37.0%	0.1%	0.1%	0.0%

Evaluating Explanation Quality

Dataset	Unconstrained inputs		Inputs with ≤ 50% difference	
	Anchor	ApproxMC3	Anchor	ApproxMC3
adult	0.99	0.67	0.99	0.81
lending	0.99	0.87	0.99	0.92
rcdv	0.99	0.75	0.99	0.80

How ***Anchor*** measures precision of $\mathcal{E}$: $prec(\mathcal{E}) = \mathbb{E}_{\mathcal{D}(\mathcal{I}' \supseteq \mathcal{E})}[\mathcal{M}(\mathcal{I}') = \pi]$

How ***we*** measure it:
approximate model counting of $\mathcal{E} \land \mathcal{M} \land \neg\pi$

Adversarial Examples

perturbed image

©Evtimov et al. CoRR abs/1707.08945

Is there a ***connection*** between **adversarial examples** and **explanations**?

Counterexample vs. Explanation Duality

Given a **classifier** $\mathcal{M}$ and **prediction** π,

a ***counterexample*** to π is subset-minimal C s.t. $C \models \bigvee_{\rho \neq \pi}(\mathcal{M} \rightarrow \rho)$

an ***explanation*** of π is subset-minimal $\mathcal{E}$ s.t. $\mathcal{E} \models (\mathcal{M} \rightarrow \pi)$

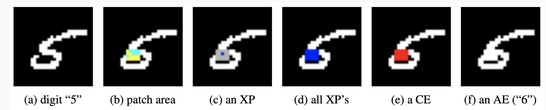
Every explanation $\mathcal{E}$ of π *breaks every* counterexample C to π
Every counterexample C to π *breaks every* explanation $\mathcal{E}$ of π

Duality-Based Enumeration

Input: formula $\mathcal{M}$ and prediction π
Output: set $\mathbb{E}$ of all explanations of π

```
(C, E, D) ← (∅, ∅, ∅)
do:
    if D ⊨ (M → π):
        E ← E ∪ {D}
    else:
        (C, ρ) ← ExtractInstance()
        for l ∈ C :
            if (C \ {l}) ⊨ (M → ρ):
                C ← C \ {l}
        C ← C ∪ {C}
    D ← MinimumHS(C)
while D ≠ ∅
return E
```

Duality Example



(a) digit "5" (b) patch area (c) an XP (d) all XP's (e) a CE (f) an AE ("6")

Figure 1: An example of digit five.

(a) digit "6" (b) patch area (c) an XP (d) all XP's (e) a CE (f) an AE ("5")

Figure 2: An example of digit six.

Acknowledgements

The author thanks his colleagues Joao Marques-Silva and Nina Narodytska, who have been taking active part in the research on rigorous logic-based XAI and coauthoring the papers, which this work extensively builds on. Without them this work would be impossible.

Follow our work on XAI:

New papers: