Explanatory Interactive Machine Learning

Stefano Teso[1], Kristian Kersting[2]

[1]KU Leuven, Belgium [2] TU Darmstadt, Germany

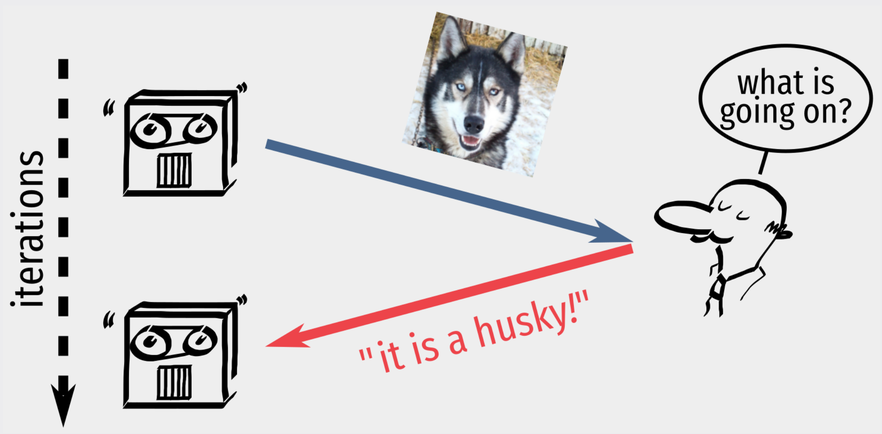
Trust ≈ Explanations + Interaction

Why explanations? Trust depends on the other's perceived "**competence**, [...], **understandability**, **directability**."

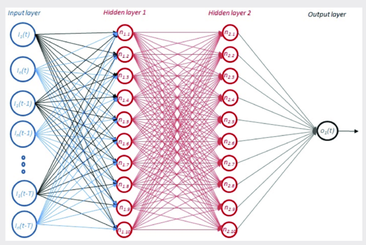
Why interaction? Trust is **dynamically updated** based on **expectations** built interactively.

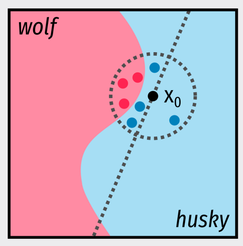
Active Learning is Opaque

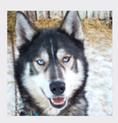
The user **a**) does not know the **model's beliefs**, **b**) can not **affect** them directly, **c**) has no clue of what his **feedback does**!

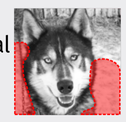


Explainable Machine Learning is Passive





x_0 =  $f(x_0)$ = *husky*

influential pixels = 

LIME [1] builds a linear approx. of model around instance x_0:
$\hat{f}(x_0) \approx \sigma\left(\sum_j w_j \phi_j(x_0) + b\right)$; $w_j =$ **influence** of feature j on $\hat{f}(x_0)$

Explainable Interactive Learning (XIL)



a) Explain predictions to user (**competence**, **understandability**),
b) Allow user to correct explanations (**directability**)

Explainable Active Learning with LIME: CAIPI

Given examples $\mathcal{L} \subseteq \mathcal{X} \times \mathcal{Y}$, instances $\mathcal{U} \subseteq \mathcal{X}$, max iters $T \in \mathbb{N}$:

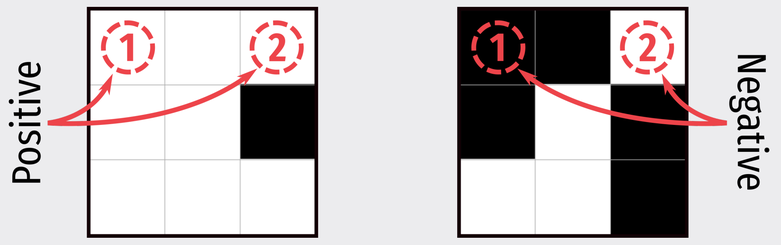
$f \leftarrow$ **Fit**$(\mathcal{L})$
repeat
  $x \leftarrow$ **SelectQuery**$(f, \mathcal{U})$
  $\hat{y} \leftarrow f(x)$, $\hat{z} \leftarrow$ **Explain**$(f, x, \hat{y})$
  Present x, $\hat{y}$, and $\hat{z}$ to the user
  Obtain y and explanation correction $\mathcal{C}$
  $\{(\bar{x}_i, \bar{y}_i)\}_{i=1}^{c} \leftarrow$ **ToCounterExamples**$(\mathcal{C})$
  Update $\mathcal{L}$ and $\mathcal{U}$
  $f \leftarrow$ **Fit**$(\mathcal{L})$
until budget T is exhausted or f is good enough

Remark: **SelectQuery**, **Fit** are underlying AL; **Explain** calls LIME; **ToCounterExamples** described below.

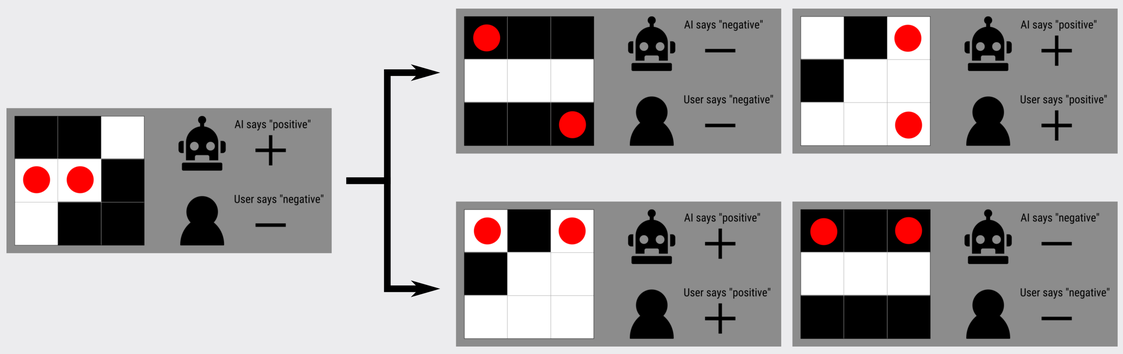
From Corrections To Counter-examples



User Study: Does CAIPI modulate trust?



We showed different interaction sessions to users:



Q1 "Do you believe that the AI eventually learned to classify images correctly?"

Q2 "Do you believe that the AI eventually learned the correct classification rule?"

Q3 "Would you like to further assess the AI by checking whether it classifies 10 random images correctly?"

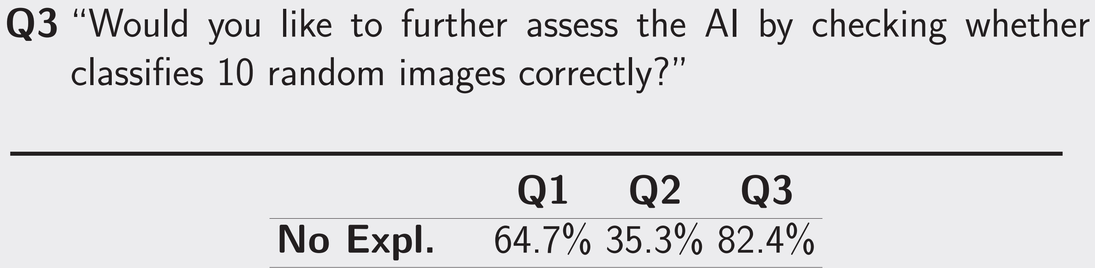
	Q1	Q2	Q3
No Expl.	64.7%	35.3%	82.4%
Bad Expl.	76.5%	64.7%	70.6%
Good Expl.	29.4%	11.8%	41.2%

Can CAIPI leverage corrections?

Accuracy vs. **input gradient constraints** [2] on Decoy Fashion MNIST (**Passive**)

	No corr.	$c=1$	$c=3$	$c=5$	IG
Train	**0.978**	0.938	0.922	0.924	0.898
Test	0.482	0.821	0.851	**0.858**	0.853

Can CAIPI learn the right concept?



L_1 SVM on the **colors** problem. **Left**: expl. F_1 score for rules 0 & 1; **Right**: decomposition of the learned weight vector;

References

[1] M. T. Ribeiro et al. Why should i trust you?: Explaining the predictions of any classifier. In *ACM SIGKDD'16*, pages 1135–1144, 2016.

[2] A. S. Ross et al. Right for the right reasons: training differentiable models by constraining their explanations. In *IJCAI*, pages 2662–2670. AAAI Press, 2017.