BEARS Make Neuro-Symbolic Models Aware of their Reasoning Shortcuts

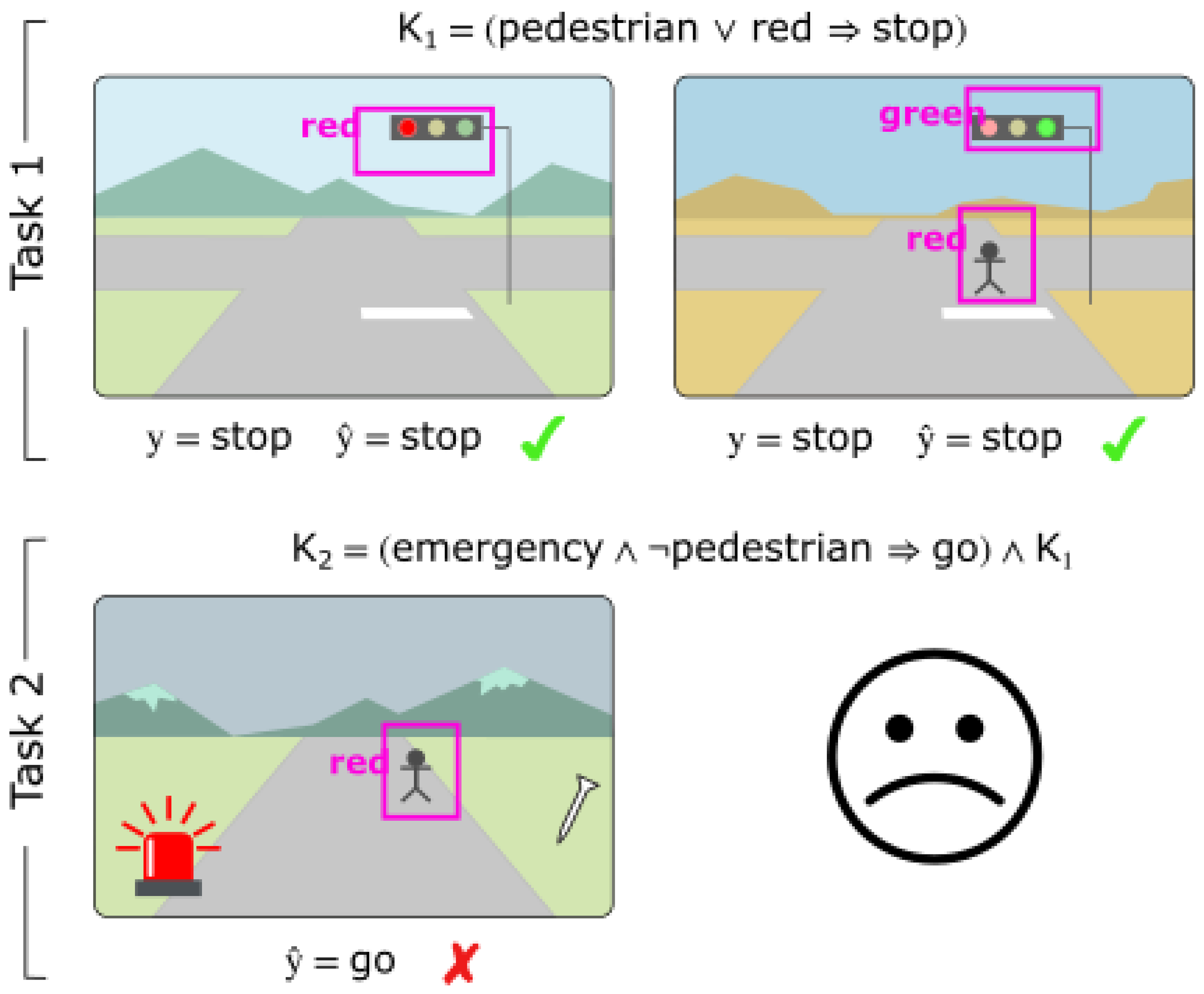
Emanuele Marconato [1,2] **Samuele Bortolotti** [1] Emile van Krieken [3]
Antonio Vergari [3] Andrea Passerini [1] Stefano Teso [1]

[1]University of Trento [2]University of Pisa [3]University of Edinburgh

REASONING SHORTCUTS

NeSy predictors such as **DeepProbLog**[1], and **Logic Tensor Networks**[2], acquire concepts that comply with the knowledge.

Are learned concepts interpretable and is the model trustworthy?
Not always![3]



Reasoning Shortcuts (RSs) like this might affect any NeSy predictor!

MITIGATION STRATEGIES

Strategy	Requires
Multi-Task	tasks
Concept Sup.	concepts
Reconstruction	(decoder)
Disentanglement	structure

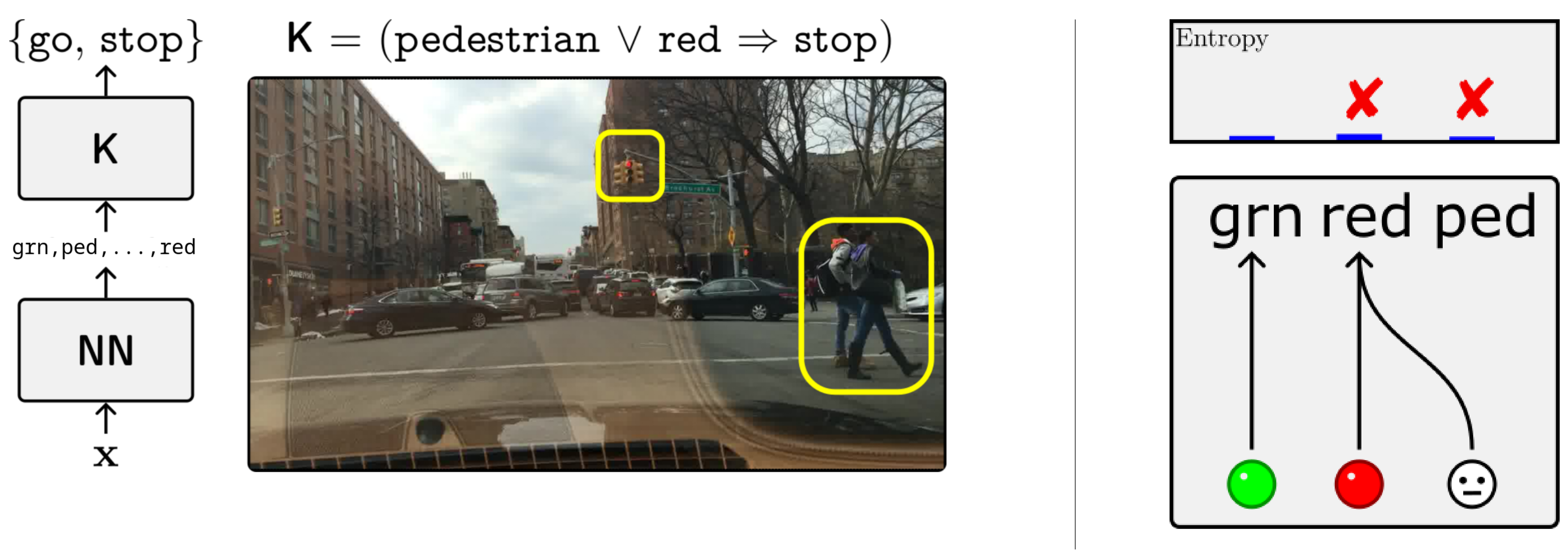
DESIDERATA

- **Concept calibration**
- **Performance**
- **Cost effectiveness**

MAIN PROBLEM

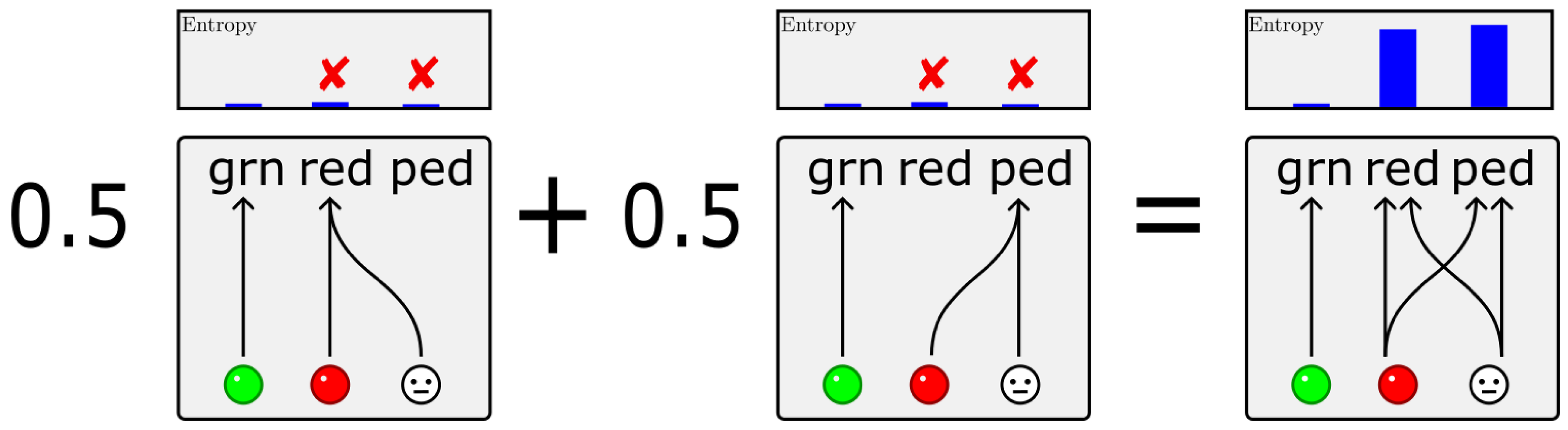
Effective mitigation strategies for RSs, like concept supervision, are often **impractical**. If the model learns a RS what concepts can we trust?

Over-confident solutions are dangerous: impossible to be aware of wrong concepts!



We propose `bears` to estimate concept uncertainty!

🐻 BEARS: BE AWARE OF REASONING SHORTCUTS! 🐻



`bears` combines **Deep Ensembles + diversification** ($\sim$ Bayesian NeSy) and provably optimizes for all desiderata:

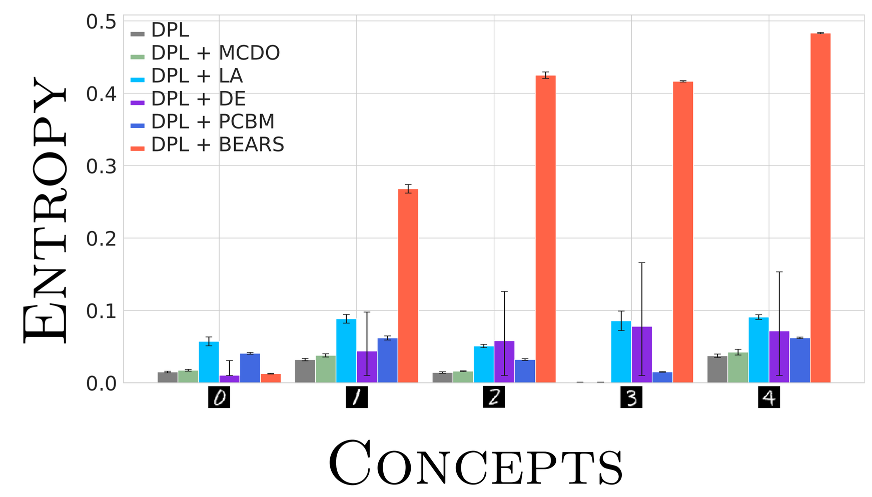
$$\mathcal{L}_{\texttt{bears}} = \mathcal{L}(\mathbf{x}, \mathbf{y}; \mathsf{K}, \theta_t)$$
$$+\gamma_1 \cdot \mathsf{KL}(p_{\theta_t}(\mathbf{C} \mid \mathbf{x}) \mid\mid \frac{1}{t}\sum_{j=1}^{t} p_{\theta_j}(\mathbf{C} \mid \mathbf{x}))$$
$$+\gamma_2 \cdot H(p_{\theta_t}(\mathbf{C} \mid \mathbf{x}))$$

EXPERIMENTS

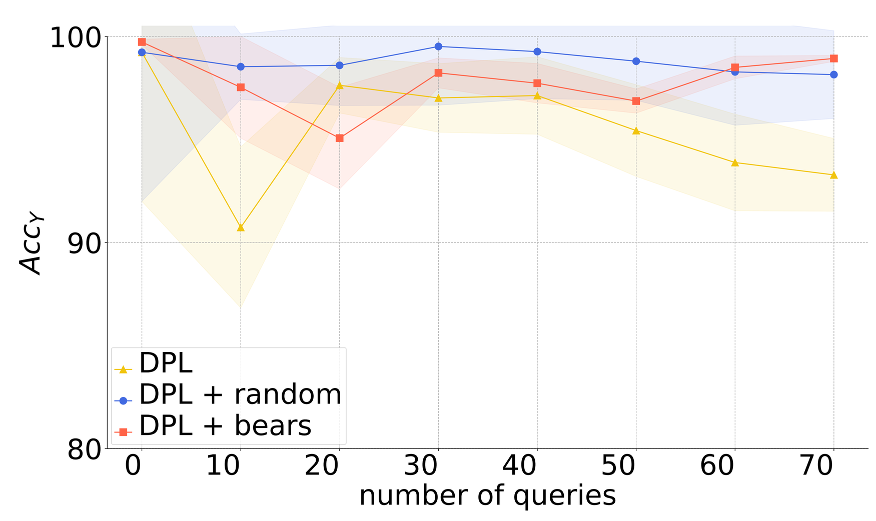
① An example from `MNIST-Addition`

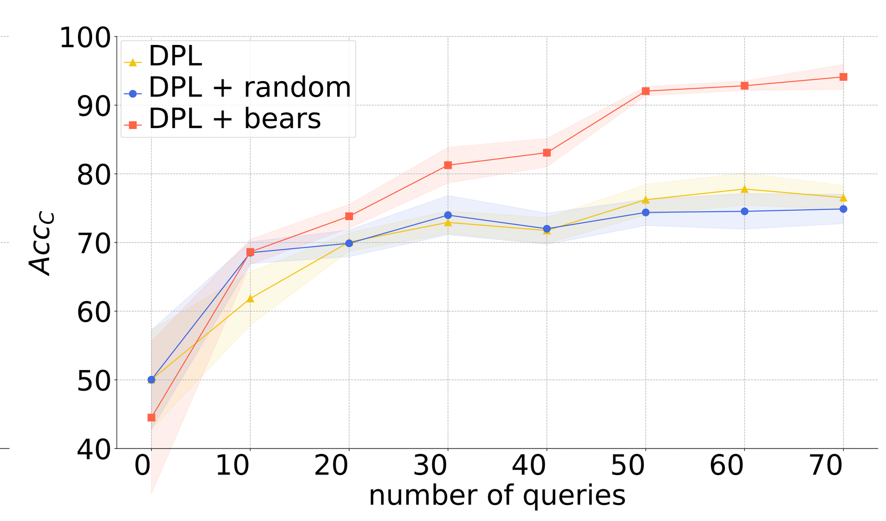
Solve the sum between two digits, *e.g.*, 2 + 3 = 5.

$$\begin{cases} 0 + 0 & = 0 \\ 0 + 1 & = 1 \\ 2 + 3 & = 5 \\ 2 + 4 & = 6 \end{cases}$$



② Active learning with `bears`





③ `bears` in real-world: BDD-OIA [4]

	mECE_C	$\mathrm{ECE}_C(F,S)$	$\mathrm{ECE}_C(R)$	$\mathrm{ECE}_C(L)$
`DPL`	0.84 ± 0.01	0.75 ± 0.17	0.79 ± 0.05	0.59 ± 0.32
+ `MCDO`	0.83 ± 0.01	0.72 ± 0.19	0.76 ± 0.08	0.55 ± 0.33
+ `LA`	0.85 ± 0.01	0.84 ± 0.10	0.87 ± 0.04	0.67 ± 0.19
+ `PCBM`	0.68 ± 0.01	0.26 ± 0.01	0.26 ± 0.02	0.11 ± 0.02
+ `DE`	0.79 ± 0.01	0.62 ± 0.03	0.71 ± 0.10	0.37 ± 0.12
+ `bears`	$\mathbf{0.58 \pm 0.01}$	$\mathbf{0.14 \pm 0.01}$	$\mathbf{0.10 \pm 0.01}$	$\mathbf{0.02 \pm 0.01}$

REFERENCES

[1] Manhaeve *et al.*, DeepProbLog, NeurIPS (2018)
[2] Donadello *et al.*, Logic Tensor Networks, IEEE (2018)
[3] Marconato et al., Not All Neuro-Symbolic Concepts are Created Equal: Analysis and Mitigation of Reasoning Shortcuts, NeurIPS (2023)
[4] Xu *et al.*, *BDD-OIA dataset*, CVPR (2020).

PAPER CODE

samuelebortolotti samuele.bortolotti@unitn.it