Shortcuts and Identifiability in Concept-based Models from a Neuro-Symbolic Lens

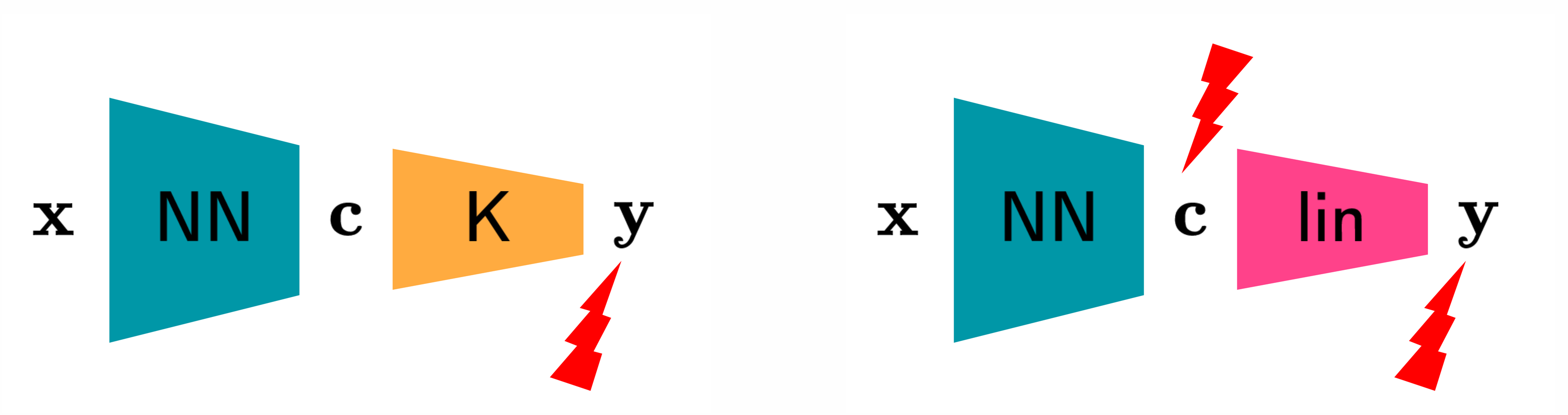
Samuele Bortolotti [1] Emanuele Marconato [1,2] Paolo Morettin [1]
Andrea Passerini [1] Stefano Teso [1]

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

CONCEPT-BASED MODELS

Concept-based models (CBMs) [1] learn a concept extractor that maps inputs to high-level **concepts** and a **linear layer** that predicts labels.

Neuro-Symbolic CBMs (NeSy-CBMs) [2] are a special case of CBMs where the linear layer is replaced with **symbolic reasoning**.



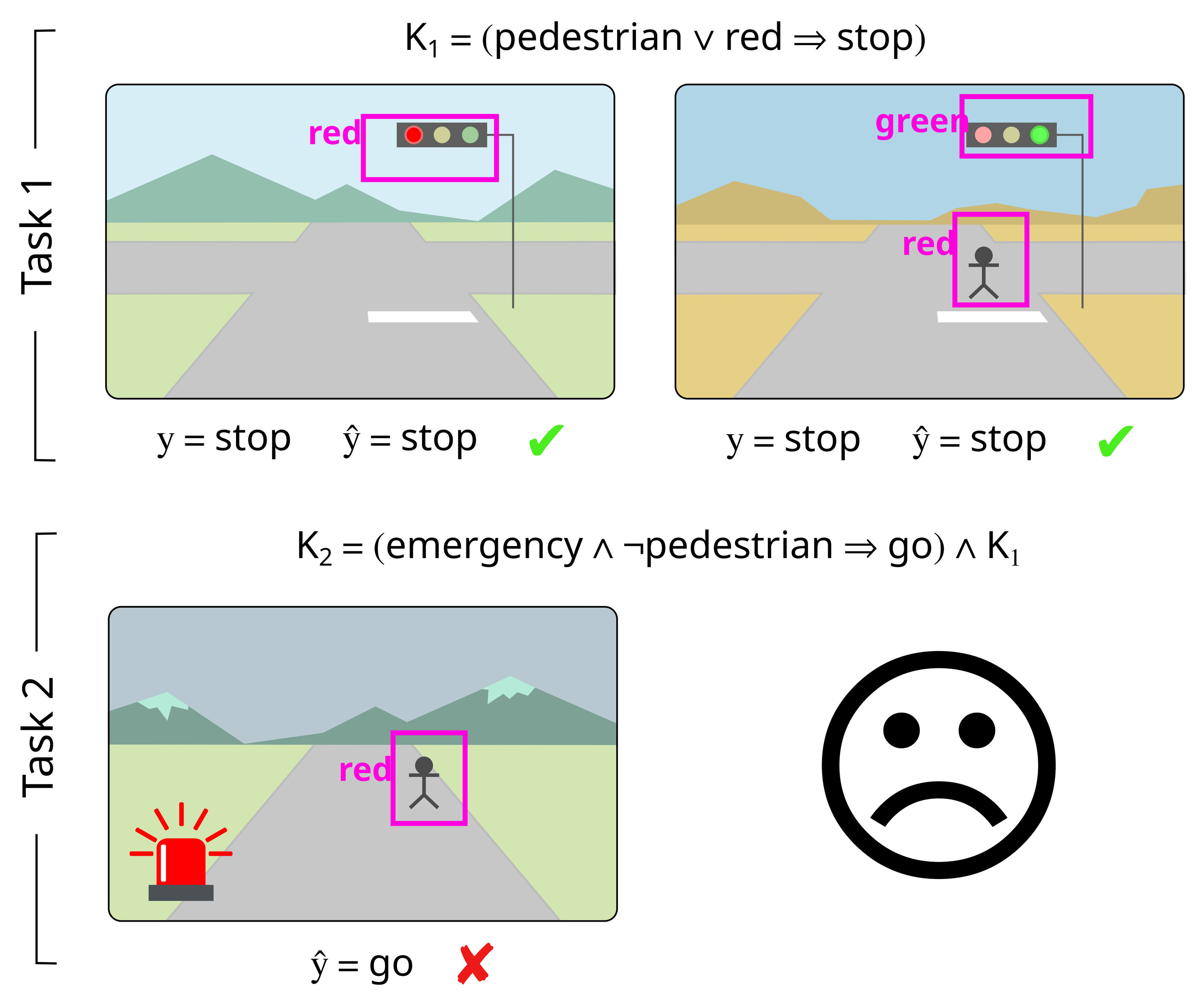
(a) NeSy-CBM (b) CBM

Do they learn correct concepts and algorithm?

INTENDED SEMANTICS

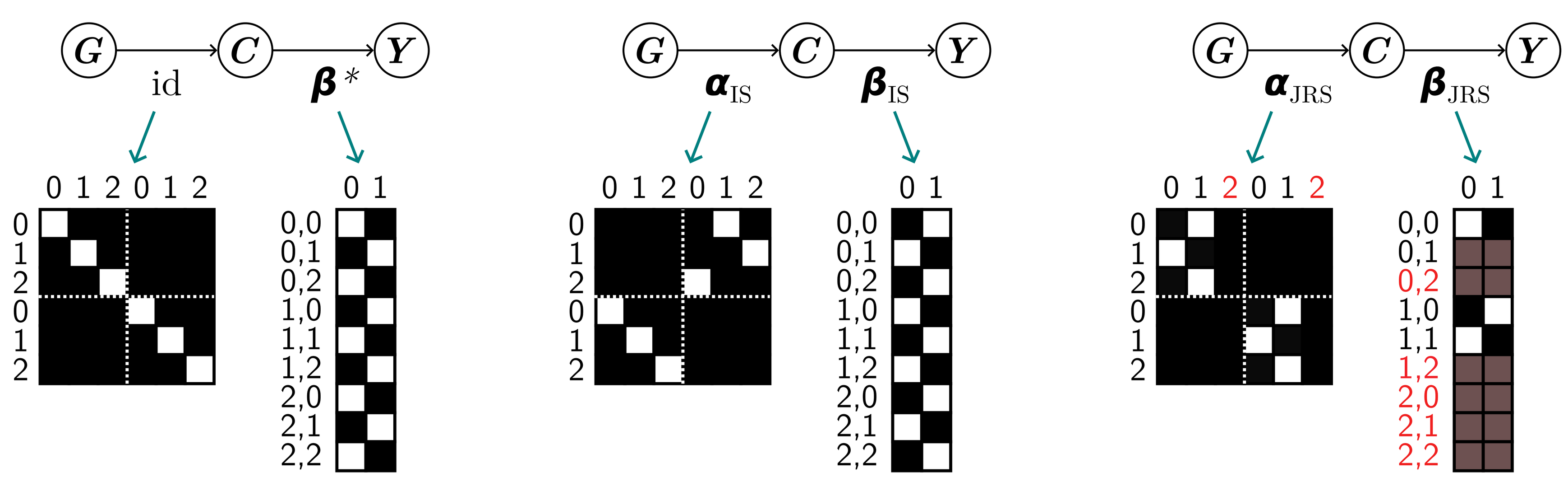
A CBM has **intended semantics** if learned concepts match ground-truth concepts and the inference layer preserves their meaning.

A well known failure case is **Reasoning Shortcuts** [3]:



What if the knowledge is learned?

`MNIST-SumParity`: find the parity of the sum between two digits, *e.g.*, (2 + 3) mod 2 = 1:



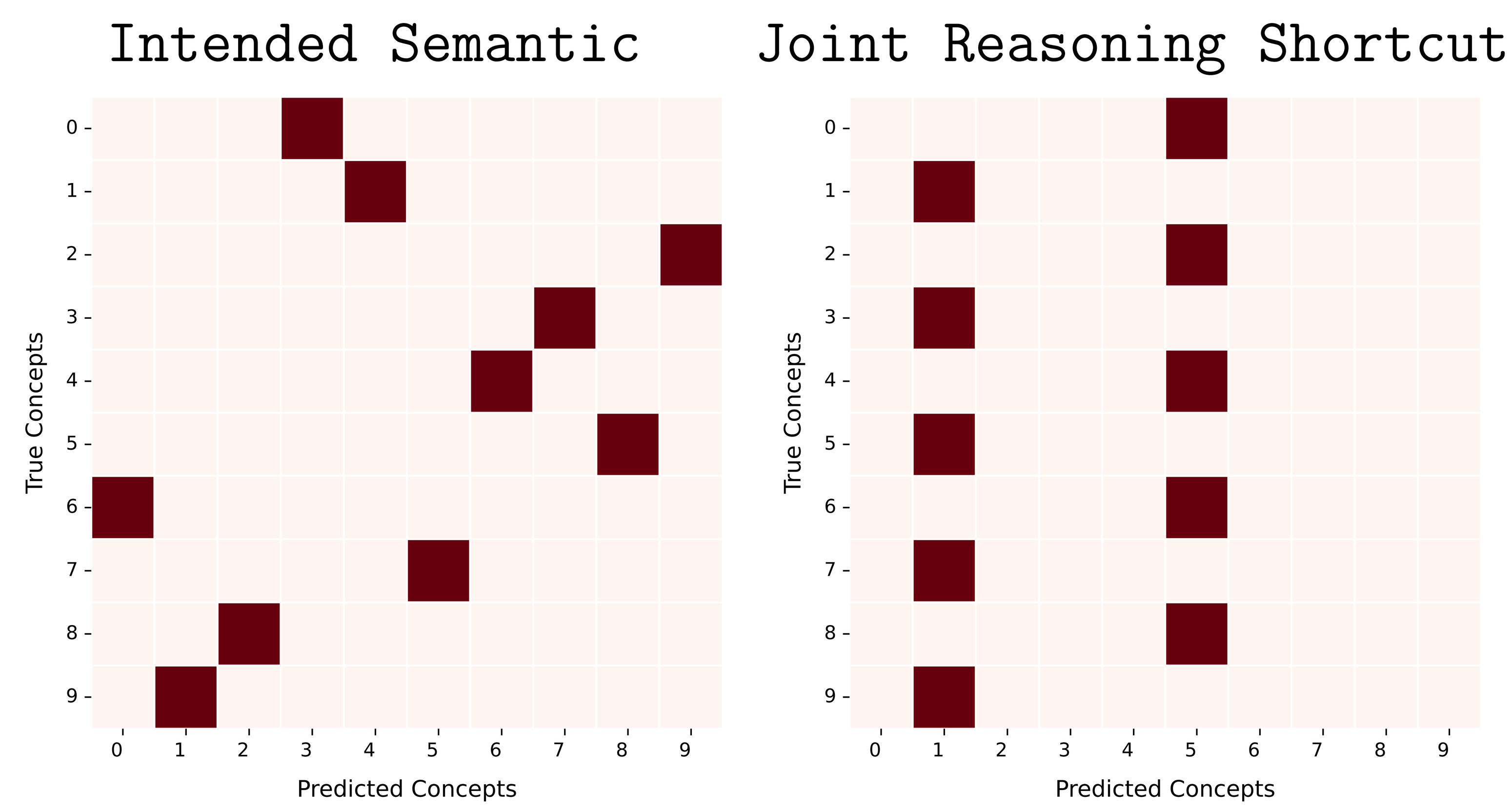
Problem: Without mitigation, the solution space is too large, making **learning the intended solution extremely challenging!**

$\sim 4.21 \times 10^6$ **unintended** solutions for `MNIST-SumParity` with digits ranging from 0 to 7.

JOINT REASONING SHORTCUTS

An example from `MNIST-SumParity`:

$y = (c_1 + c_2) \bmod 2$ $y = (c_1 - c_2)$



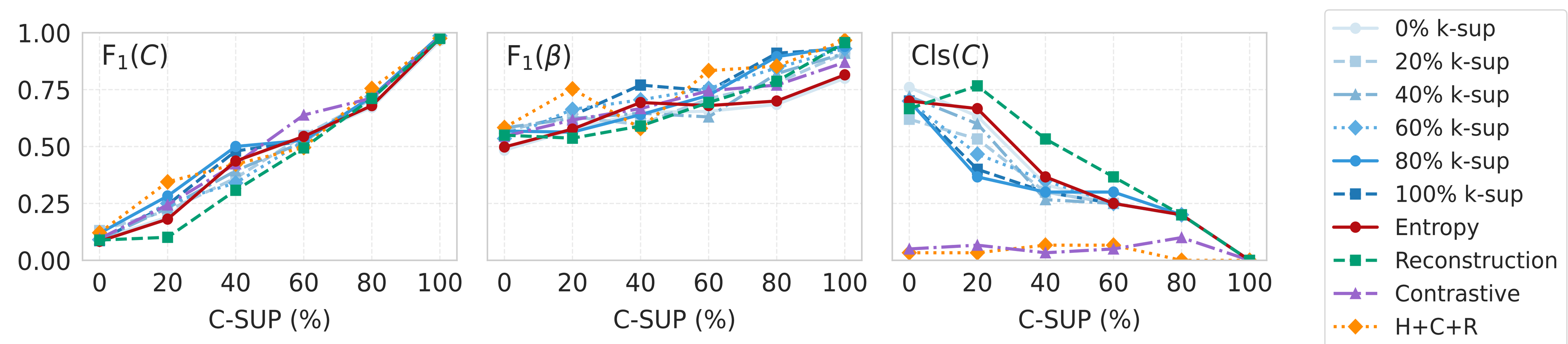
Theorem: no deterministic JRSs $\Longrightarrow$ intended semantic

MITIGATIONS

- **Supervised**: concept-sup., multi-tasks, knowledge distillation
- **Unsupervised**: disentanglement, reconstruction, contrastive

CASE STUDIES

① **Traditional mitigation strategies on `MNIST-SumParity`:**



② **Out of distribution impact on `MNIST-SumParity`:**

		ID		OOD	
C-Sup	K-Sup	$F_1(Y)$ (↑)	$F_1(\boldsymbol{\beta})$ (↑)	$F_1(Y)$ (↑)	$F_1(\boldsymbol{\beta})$ (↑)
0%	0%	0.99 ± 0.01	0.55 ± 0.05	0.01 ± 0.01	0.47 ± 0.07
0%	100%	0.99 ± 0.01	0.56 ± 0.06	0.01 ± 0.01	0.58 ± 0.08
100%	0%	0.97 ± 0.01	0.88 ± 0.04	0.40 ± 0.14	0.31 ± 0.12
100%	100%	0.97 ± 0.01	0.95 ± 0.01	0.97 ± 0.02	0.69 ± 0.27

③ **Future work**: what happens in standard networks and LLMs?

REFERENCES

[1] Koh *et al.*, Concept bottleneck models, ICML (2020)
[2] Manhaeve *et al.*, DeepProbLog, NeurIPS (2018)
[3] Marconato *et al.*, Not All Neuro-Symbolic Concepts are Created Equal: Analysis and Mitigation of Reasoning Shortcuts, NeurIPS (2023)

PAPER

samuelebortolotti samuele.bortolotti@unitn.it