Neuro-Symbolic AI

Vaishak Belle, ***University of Edinburgh***

Agenda

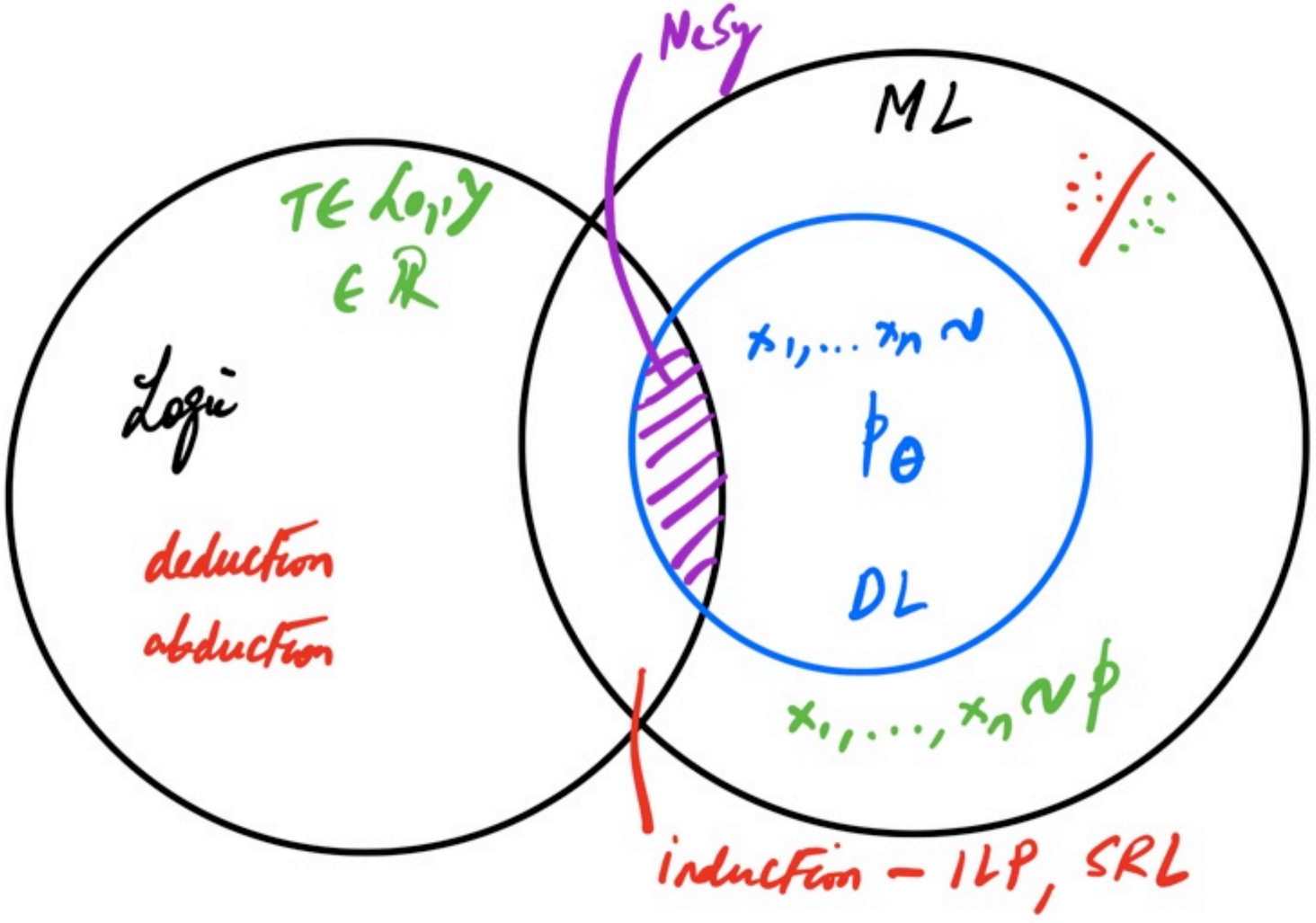
- Motivations for neuro-symbolic AI
- Logic for domain *constraints*
  - what is the hypothesis space?
  - what is the distribution space?
- Nesy program induction
- Nesy + RL
- NeSy + LLMs
- Conclusions

Challenges with Neural networks

- **Structured Reasoning:** struggle with hierarchical reasoning and causality vs. correlation
- **Data Need:** require large datasets for robust predictions
- **Knowledge Integration:** Integrating expert and common sense knowledge is challenging

- **Explainability**: raising ethical, security and HCI concerns
- **Guarantees:** output is based on highest probability, so may violate constraints in safety-critical applications

Marcus (2018)


NeSy
ML
TE Log, y
∈ ℝ
Logic
deduction
abduction
x1, ... xn ~
pθ
DL
x1, ..., xn ~ p
induction – ILP, SRL

Language	**Uncertainty**	**Training**	**Model**
Propositional/ Prolog	Probabilistic	Probabilistic	SRL
Propositional/ Prolog	Labelled examples	Entailment	ILP
Propositional	*Fuzzy*	*Deep learning*	*NeSy (LTN)*
Propositional	*Probabilistic*	*Deep learning*	*NeSy (ProbLog)*

A few exciting and emerging areas

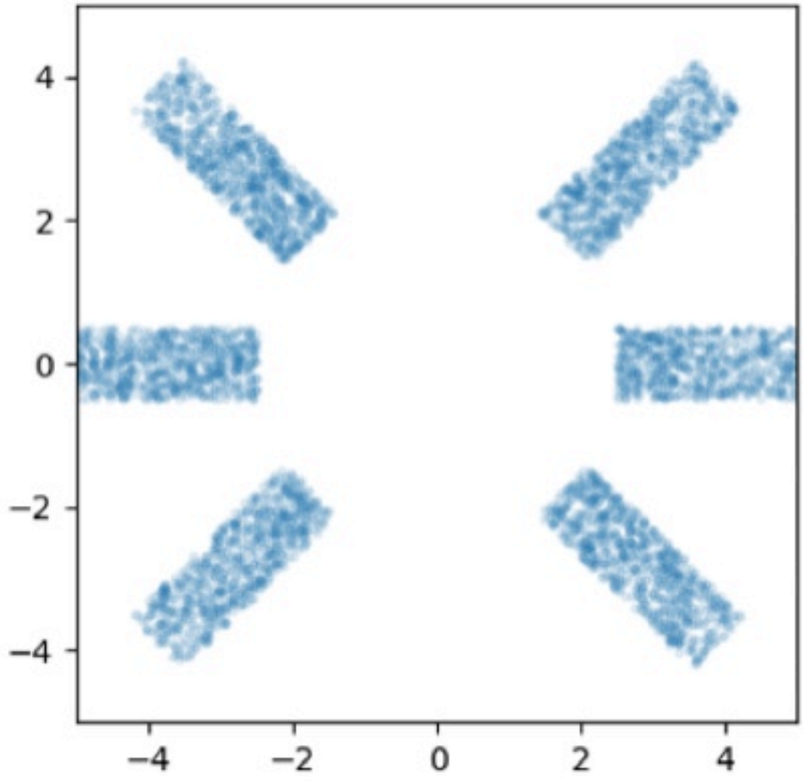
- **Knowledge graphs:** knowledge completion, knowledge-based reasoning
- **Neuro-symbolic logic programs:** hybrid reasoning with logical and neural predicates
- **differential inductive programming:** program and structure induction using neural techniques

- **logic-based loss functions:** enforce geometric and logical constraints during predictions
- **Nesy + RL:** human-level logical feedback to train dynamic systems
- **LLMs:** as a service for constructing knowledge or to improve answers involving reasoning

Logic for domain constraints

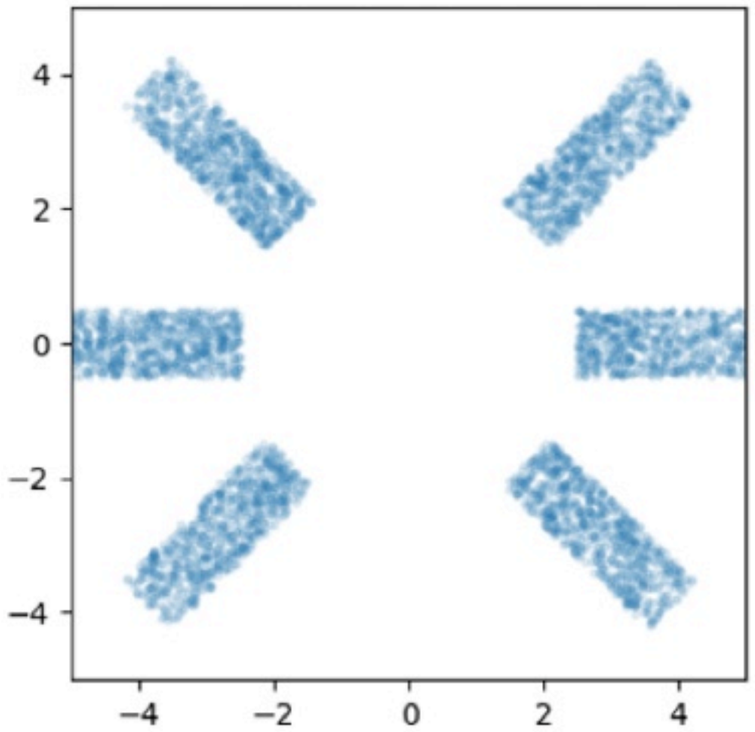
neural network regularisation using logical formulas

Density estimation



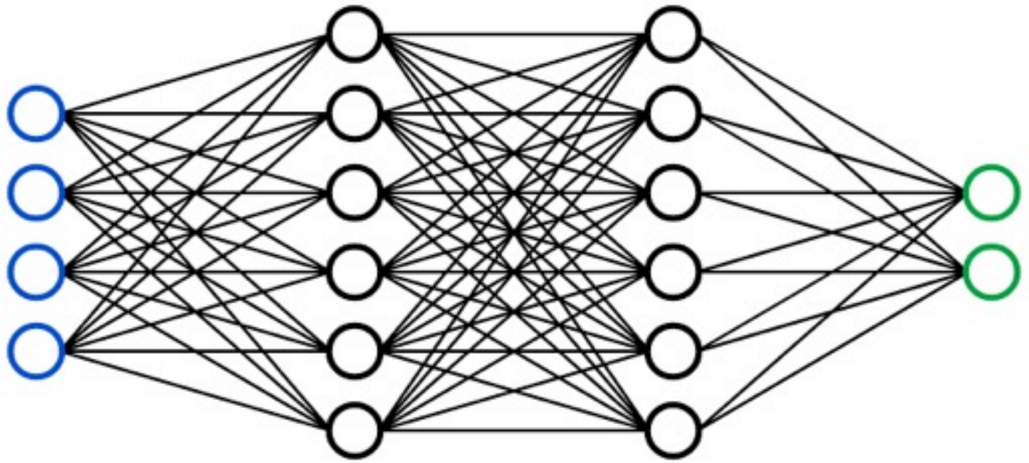
$$\Phi = (x_1 > -.5 \wedge x_1 < .5 \wedge x_2 > .5 \wedge x_2 < 4) \vee \ldots$$

$$\vee (x_1 + x_2 > -.5 \wedge x_1 + x_2 < .5 \wedge x_1 - x_2 > .5 \wedge x_1 - x_2 < 4)$$



Formulation

$$X = \{x^{(0)}, \ldots, x^{(N)} \mid x^{(i)} \sim^{iid} p^*(x), x^{(i)} \models \Phi\}$$



$$p_\theta(x)$$

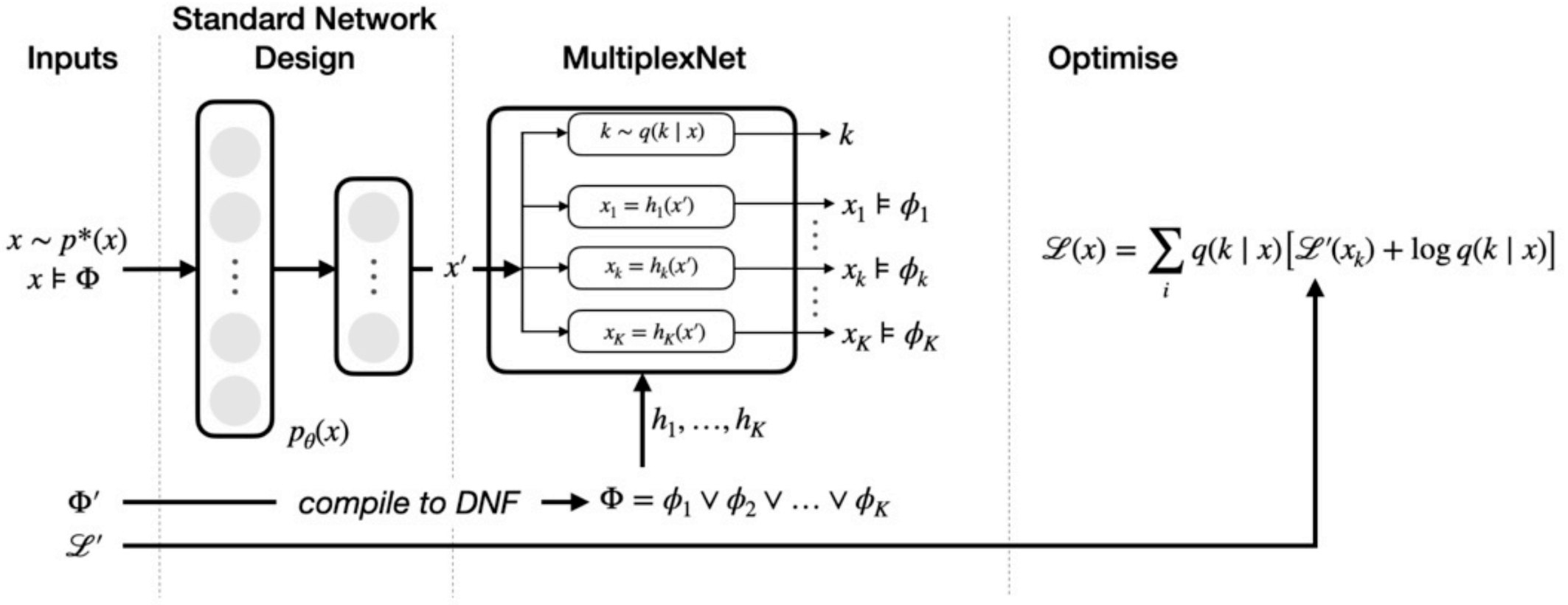
Model

Design: $p_\theta(x)$

Train: $p_{\theta^*}(x) = \arg\max_\theta(\log p_\theta(X))$

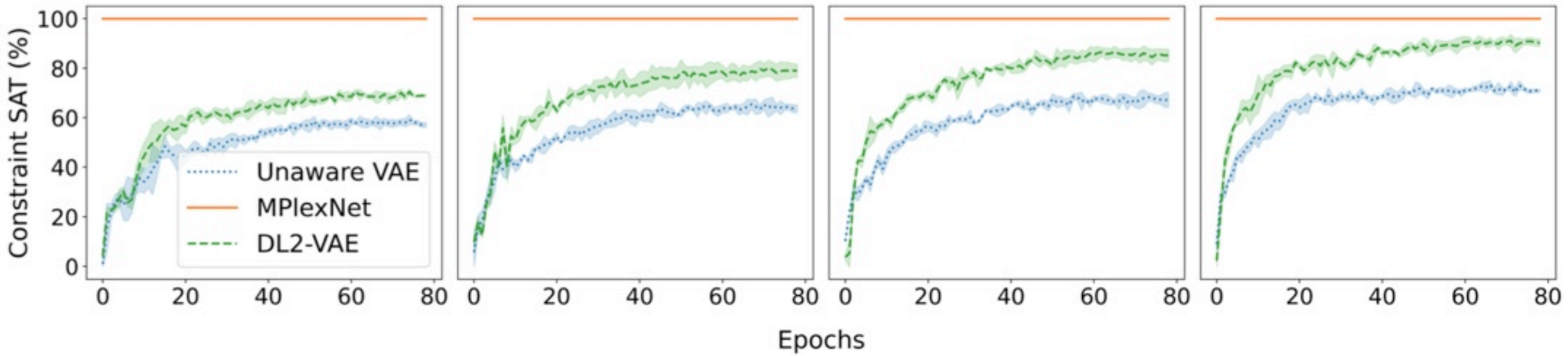
But *constraint*?? $+ L_\Phi(X)$

Pipeline

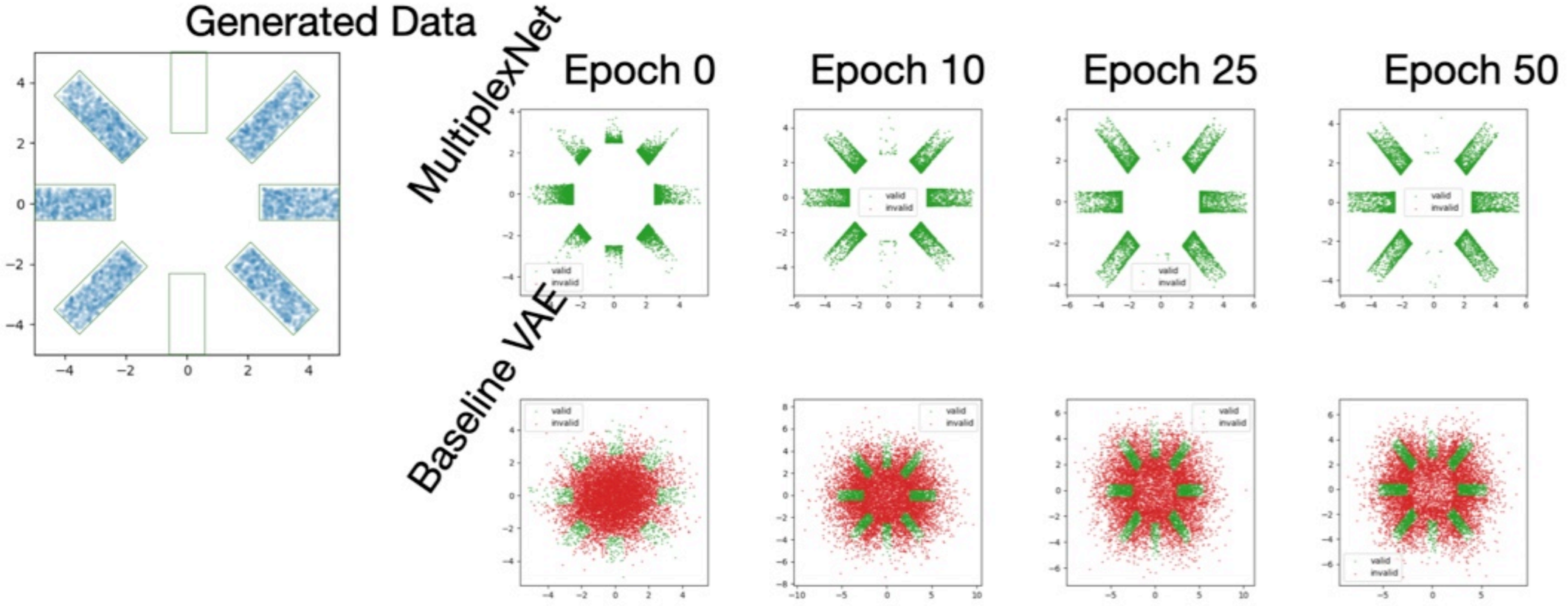


Hoernle et al. 2022; cf. semantic loss, DeepProbLog, DL2

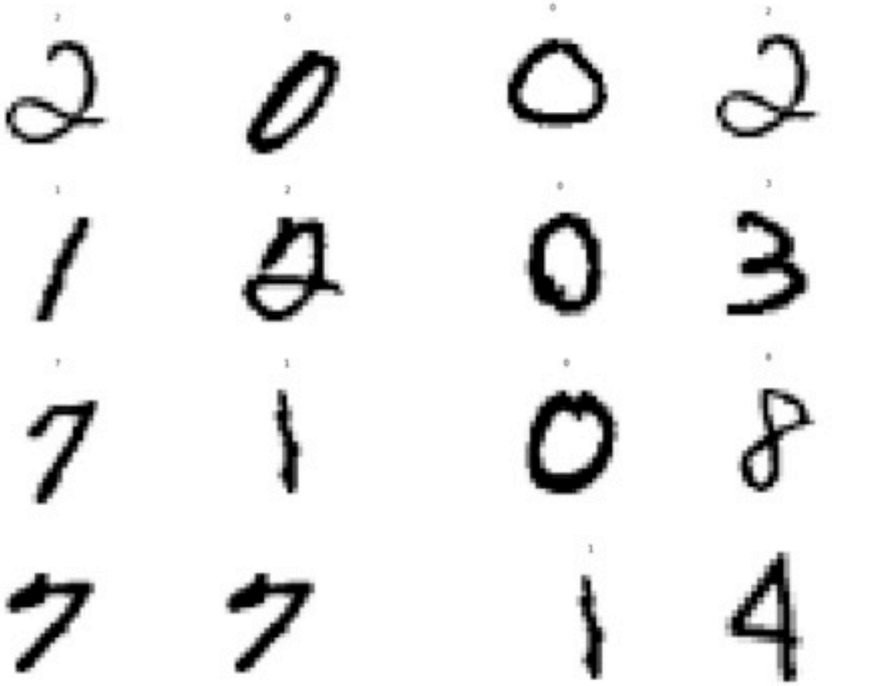
Constraint satisfaction



Vs Baseline VAE



Delayed signals = weak supervision



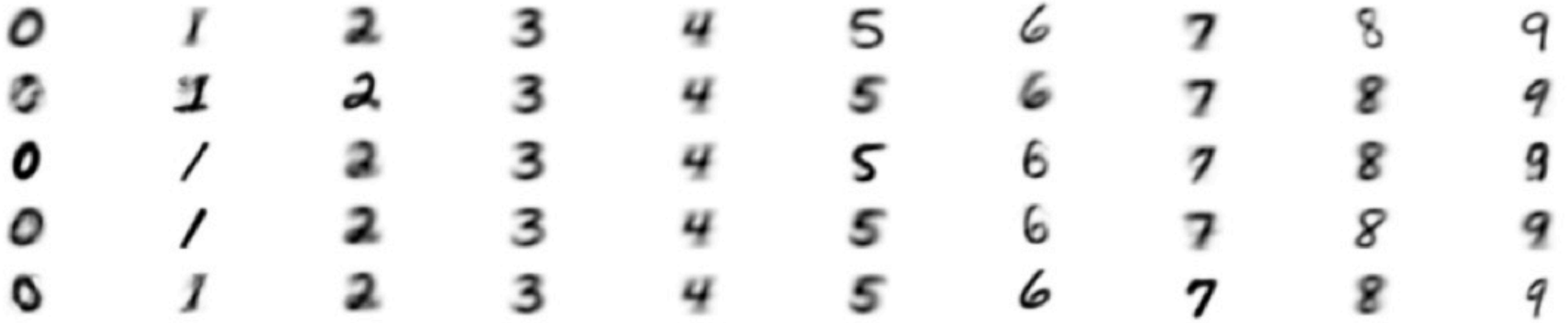
Constraint

$$\Phi = (y_1 = 0 \land y_2 = 0 \land y_3 = 0 \land y_4 = 0) \lor \dots$$

$$(y_1 = 0 \land y_2 = 1 \land y_3 = 0 \land y_4 = 1) \lor \dots$$

$$(y_1 = 9 \land y_2 = 9 \land y_3 = 1 \land y_4 = 8)$$

Decoded samples



Super-class with CIFAR100 (e.g., maple is also tree)

Model	Class Accuracy	Super-class Accuracy	Constraint Satisfaction
Vanilla ResNet	$75.0 \pm (0.1)$	$84.0 \pm (0.2)$	$83.8 \pm (0.1)$
Vanilla ResNet (SC only)	NA	$83.2 \pm (0.2)$	NA
Hierarchical Model	$71.2 \pm (0.2)$	$84.7 \pm (0.1)$	$\mathbf{100.0 \pm (0.0)}$
DL2	$\mathbf{75.3 \pm (0.1)}$	$84.3 \pm (0.1)$	$85.8 \pm (0.2)$
MultiplexNet	$74.4 \pm (0.2)$	$\mathbf{85.4 \pm (0.3)}$	$\mathbf{100.0 \pm (0.0)}$

Hierarchical is trained to predict the super class label and class label, conditioned on the value for the super class label; represents the bespoke engineering solution

Characterising the learned model

Towards correctness / verifiability of the hypothesis space

Wang et al 2024, Uperti & Belle 2024 (ongoing)

Setup

Input: $(x_1, x_2, \ldots, x_n)$

Classifier output: $\in \{y_1, y_2, \ldots, y_k\}$

Signal: $\sigma(\vec{y})$!

E.g., $\sigma \in \{+, - \times \oplus\}$

Goal: *Attempt an ILP formulation!*

Logical predicates

- $CP(f_i, x_i, y_i) \equiv f_i(x_i) = y_i$
- $TP(\vec{y}, s) \equiv \sigma(\vec{y}) = s$
- $OP(\vec{x}, s) \equiv D(\vec{x}) = s$

Given $(OP(\vec{x}, s) \wedge \bigwedge_i CP(f_i, x_i, y_i))$, and (known) σ s.t.

$TP(\vec{y}, s)$.

ILP formulation

Examples *E*:

$$\{(\vec{x}, s) | OP(\vec{x}, s) \wedge \bigwedge_i CP(f_i, x_i, y_i) \wedge TP(\vec{y}, s) \wedge \text{label}(\sigma(\vec{y})) = s\}$$

Find H s.t., $B \wedge H \models E$.

Where B is (say) axioms about numbers

Example with known σ (=+)

Given: label(+($\underline{0}$, $\underline{1}$)) = 1, label(+($\underline{2}$, $\underline{3}$)) = 5, etc.

Find *H* (extension to label) s.t. $B \wedge H$ entails:

$$\exists x_0, x_1, x_2, x_3, \ldots \mathrm{label}(\underline{0}) = x_0 \wedge \mathrm{label}(\underline{1}) = x_1 \wedge \ldots$$

$$\mathrm{label}(x_0 + x_1) = 1 \wedge \mathrm{label}(x_2 + x_3) = 5 \ldots$$

Thus, $H \doteq \mathrm{label}(\underline{0}) = 0, \ldots$

Variations

- Permits k-ary functions
- Guess σ as well as H?
- How to deal with (infinitely) many variations of images?

A geometric view on correctness

Baseline trained distribution: q

Constrained distribution: p ... both drawn from a family of distributions $\mathcal{D}$.

Task: *minimize* $d_{\mathcal{D}}(p, q)$

E.g., Kullback-Leibler $d_{KL}(p||q) \doteq \sum_x p(x) \log \frac{p(x)}{q(x)}$.

Classifier viewpoint

Loss: $L_\alpha(f) \doteq d_{\mathcal{D}}(p_\alpha, f)$.

Constrained distribution: $p_\alpha(x) = \mathcal{U}(x \in \mathcal{M}(\alpha))$.

or some other on $\mathcal{M}(\alpha), \ldots$

So includes WMC-based semantic loss, SAT, etc.

Error minimization (for XOR propositional constraint)

Init. Dist.(approx)	[1 0 0 0]	[0 1 0 0]	[.5 0 0 .5]	[.33 .33 .33 0]	[.25 .25 .25 .25]
L^2-Norm	0.01201	2.74626	1.29408	0.00113	0.00130
FisherD.(ours)	**8.821e-06**	**1.782e-05**	**2.444e-06**	**1.889e-05**	**5.364e-06**
KL-Div(ours)	0.00097	0.00046	0.00102	0.00078	0.00082
-WMC	0.82146	4.01373	0.52928	0.02228	0.00114
Sloss	3.48912	4.01371	0.38087	0.02220	0.00111

Neural ILP

- combine deep learning with relational ILP
- vs deep learning: explanatory power
- vs ILP: deal with **noisy, continuous, non-linear** data

Evans & Grefenstette, 2017; Payani & Fekri, 2019; Bueff & Belle, 2024

E.g., given data that generates $e = mc^2$, can you learn such a rule/equation?

3 steps

- continuous data discretisation
- operational predicates like addition or multiplication or square or root
- rules that combine these

Remark: classifier based, so rules observed for intervals of relevant variables

Example

Recall: $B \wedge H \models \gamma, \forall \gamma \in P$, and $B \wedge H \not\models \gamma, \forall \gamma \in N$

- B = {Car(ford), Clothing(jacket), On(jacket, bob), Inside(carol, volvo), ⋯}
- P = {Passenger(bob), Passenger(carol), ⋯}
- N = {Passenger(volvo), Passenger(jacket), ⋯}

$$\texttt{Passenger}(X) \leftarrow \texttt{Inside}(X, Y1) \wedge \texttt{Car}(Y1) \wedge \texttt{On}(Y2, X) \wedge \texttt{Clothing}(Y2)$$

If an object is inside the car with clothing on it, then it is a passenger

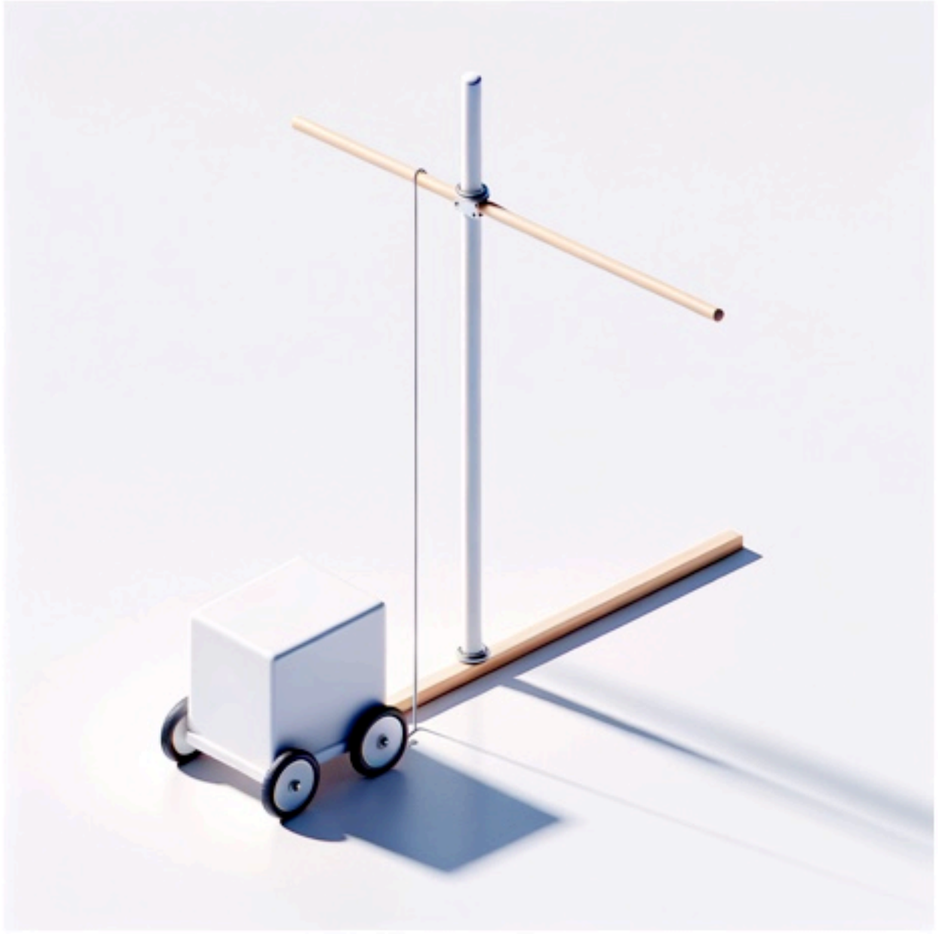
In our setting, we might have ...

$$\texttt{Class1}(m, c)_{(0 \leq E \leq 3.07 \times 10^{-1})} \leftarrow$$

$$\texttt{LessThan}(m, 0.6) \wedge \texttt{SquareLessThan}(c, 0.5) \wedge \texttt{ProdLessThan}(m, c2, 0.3)$$

if m is less than 0.6 and the square of c is less than 0.5 and the product between m and square of c is less than 0.3, then the value of E will be greater than or equal to 0 and less than or equal to 3.07×10^{-1}

Cart pole (RL)



CartPole	Policy rules for dNLRLnlc
mean reward: 294.7 ± 25.8	**left**()
	$: -([0.60]CartPos < 2.57 \wedge PoleAngleSine > 0.00 \wedge PoleAngleVeloc > -0.38)$
	right()
	$: -(PoleAngleSine < 0.04 \wedge PoleAngleVeloc < 0.00)$
	$: -(CartPos < 0.74 \wedge [0.55]CartVeloc > -1.64 \wedge CartVeloc < -0.11$
	$\wedge PoleAngleSine < 0.65 \wedge [0.66]PoleAngleVeloc > -2.04 \wedge PoleAngleVeloc < 0.28)$

Bueff and Belle, 2023.

Using LLMs as "encoder" for language

Premise

There are four persons. Everyone is visible to others. Each person draws a card, face unrevealed (red or black). Cara's card is shown to Vasiliki. Cara's card is shown to Conrad. Jennifer's card is shown to Conrad. Vasiliki's card is shown to Cara. It is publicly announced that someone picked a red card. It is publicly announced that Vasiliki knows whether someone picked a red card.

Hypothesis

Cara can now know whether Conrad picked a red card.

Symbolic Formulation

VARS 1,2,3,4

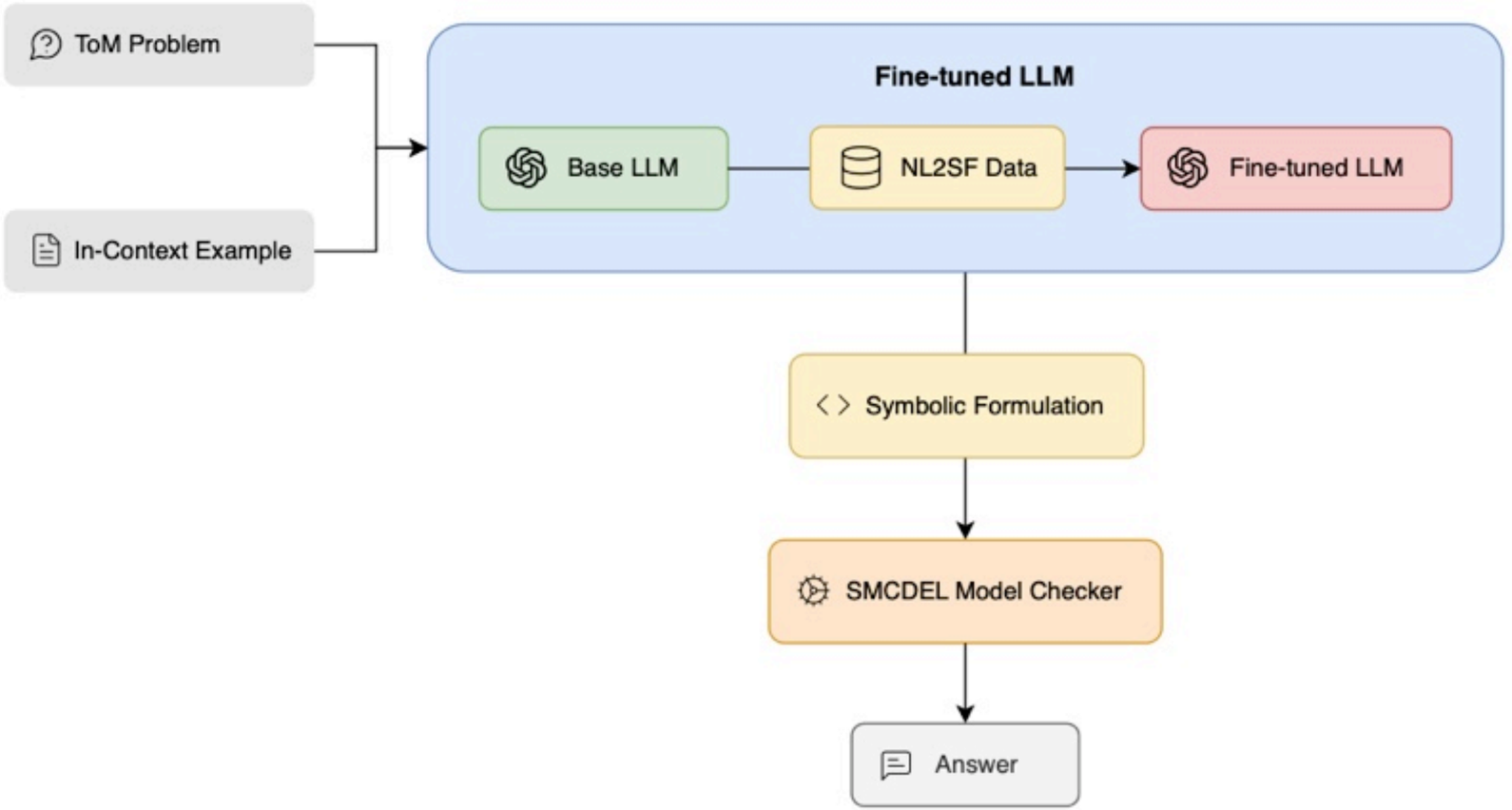
LAW Top

OBS Agenta:3 Agentb:3,4 Agentc:1

VALID? [! (1|2|3|4)] [! (Agenta knows whether (1|2|3|4))] (Agentc knows whether 2)

Sileo and Lernould, 2023; Tang and Belle, 2024

Pipeline


ToM Problem
In-Context Example
Fine-tuned LLM
Base LLM
NL2SF Data
Fine-tuned LLM
Symbolic Formulation
SMCDEL Model Checker
Answer

On Mindgames (DEL + public announcement)

Approach	Execution Rate(%)	Accuracy(%)	AUC
DP	99.50	58.00	0.58
SFGP	78.00	49.00	0.60
DP_{FT}	**100**	76.00	0.76
ToM-LM	94.50	**91.00**	**0.94**

Symbolic reward shaping

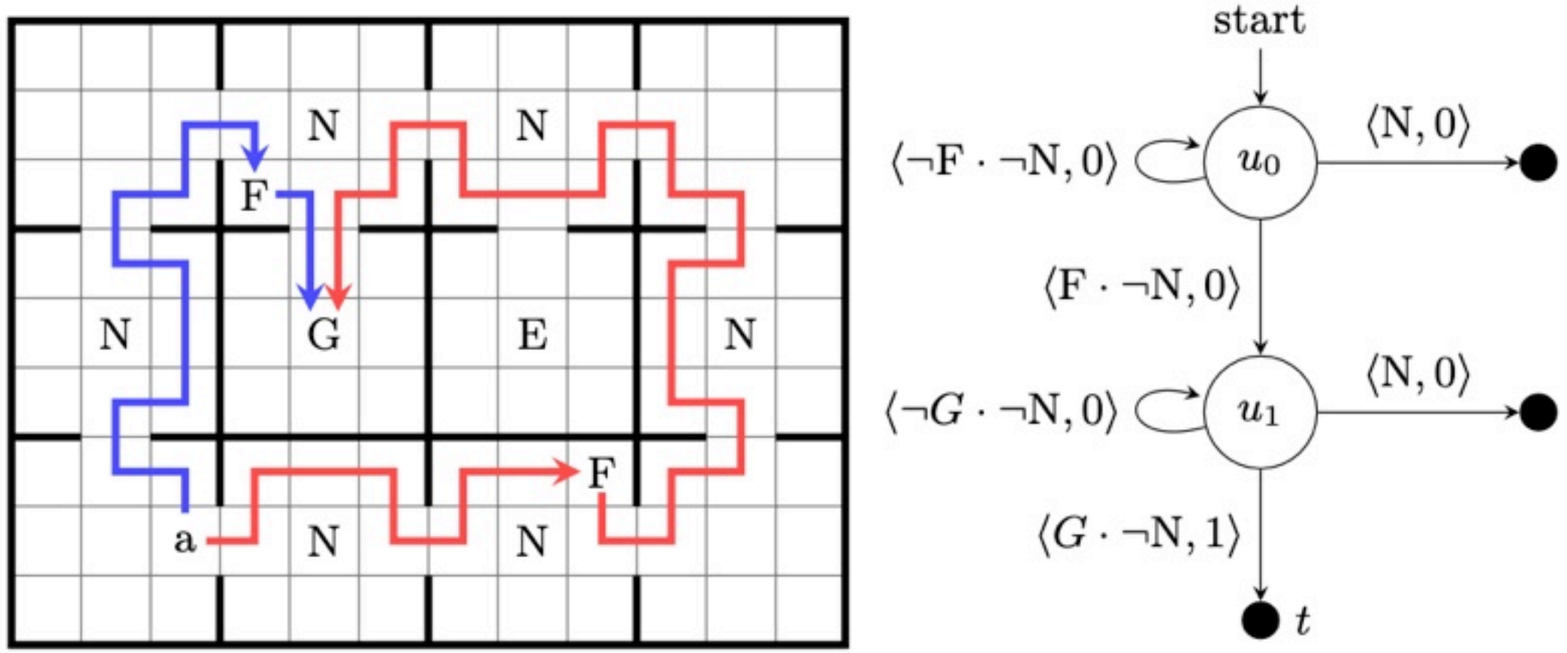
Recall: MDP is tuple

$$M = \langle S, A, R, P, \gamma \rangle; R : S \times A \to \mathbb{R}; P : S \times A \times S \to \mathbb{R}_{[0,1]}; \gamma \in (0, 1]$$

Suppose *R* were specified by a logical formula?

Icarte et al. (2022)

Driving world



(a) Driving World Environment

(b) Single Task ERM

Ciupa and Belle (2024)

Moral labels

Suppose we were able to label **good, bad** and **neutral** actions.

A plan $\pi = a_0 \ldots a_n$ is **morally permissible** according to Act-Deontology if and only if $\neg Bad(a_i), \forall i$ holds.

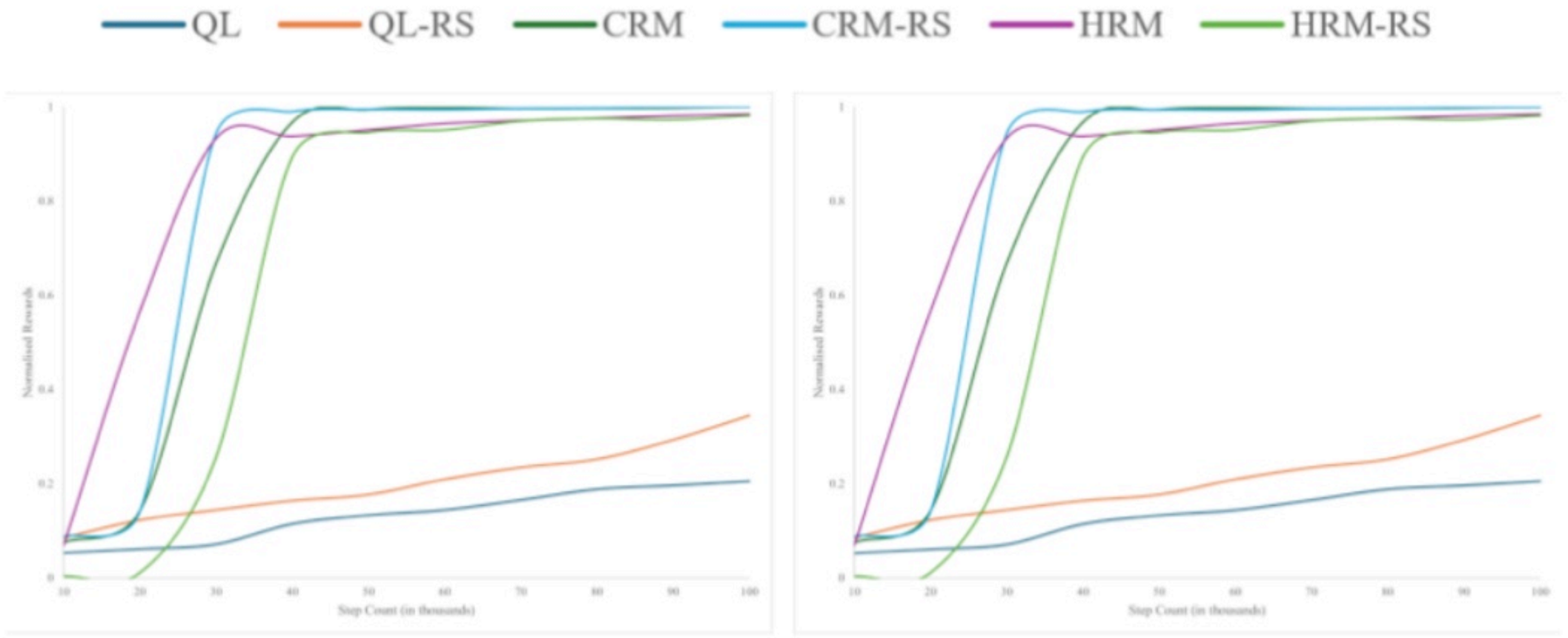
Looked at **driving, trolley dilemma** and learning from **moral preferences**

Ciupa and Belle (2024), Ciupa and Belle (ongoing)

Algorithm & evaluations

Algorithm 1 Reward Machines with Act Deontology

Input: MDPRMs, Π
Output: Policy for MDPRMs $\times \Pi'$ ($\Pi' = (\Pi \times$ Act-Deontology$)$) or Impermissible
episode_ended $\leftarrow$ **False**
while $\neg$*episode_ended* **do**
    Observe s' from a'
    Check moral result of $\forall s \in S$, $s' = \{\text{Good}(s) \vee \text{Neutral}(s) \vee \text{Bad}(s)\}$
    if s' is Good$(s) \vee$ Neutral(s) **then**
        episode_ended $\leftarrow$ **False**
        Policy for MDPRMs $\times \Pi'$
    else
        episode_ended $\leftarrow$ **True**
        Impermissible
    end if
end while



(a) Act Deontology

(b) Utilitarian

Summary

- Logic for *loss functions*
- Logic with *program induction + LLMs + RL*

Logic for ... constraints, explanations, correctness

Outlook

- Neuro-symbolic AI is a rich landscape covering various strategies and frameworks for integrating deep learning and reasoning
- Often referred to as the third wave of AI, combining the best of both worlds

- Not reliant on a single foundation, allowing for ad-hoc constructions -- presents challenges from a foundational standpoint
- The most promising direction for knowledge integration and correctness/verification