EXPLAINABLE ARTIFICIAL INTELLIGENCE

L193 – Lecture 6 - Lent 2025

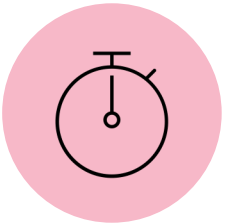
UNIVERSITY OF CAMBRIDGE

NEURAL INTERPRETABLE REASONING

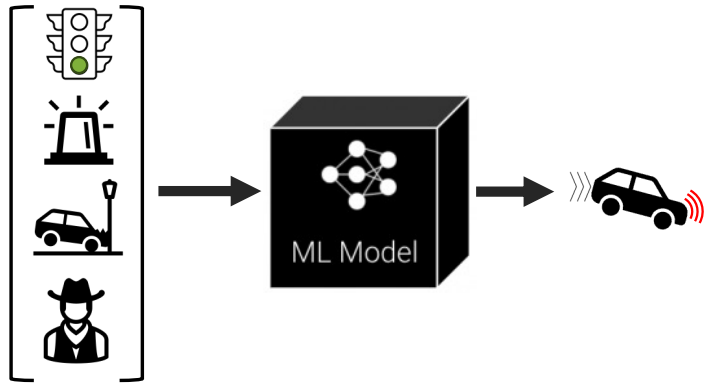
Pietro Barbiero, PhD

IBM **Research**

THE COMPLEXITY OF UNDERSTANDING



Are <20 interactions sufficient to fully understand this ML model?

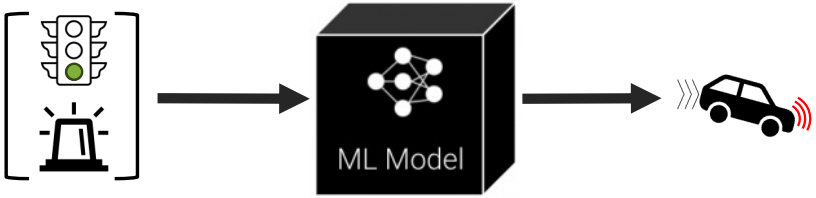


If yes, raise your hand!

Disclaimer: Unfortunately for you, I'm always right during the lecture—but I'm happy to be proven wrong afterward! ☺

THE COMPLEXITY OF UNDERSTANDING

How many interactions do we need to fully understand the model?

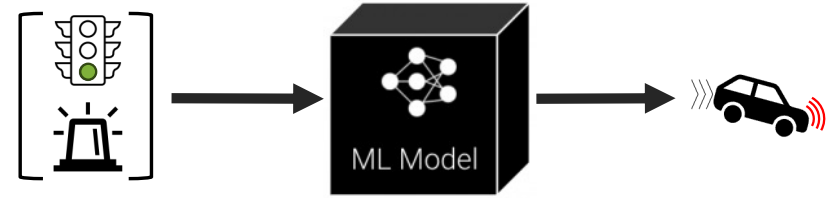


$c_i \in \{0,1\}$

0	0	1

THE COMPLEXITY OF UNDERSTANDING

How many interactions do we need to fully understand the model?



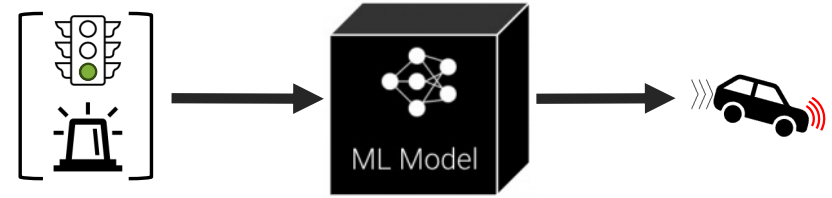
$c_i \in \{0,1\}$

0	0	1
1	0	0

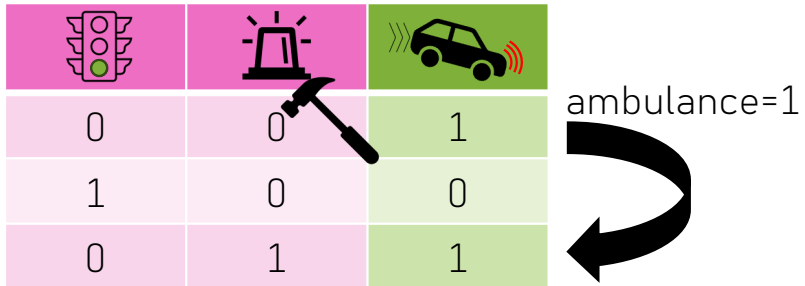
green light=1

THE COMPLEXITY OF UNDERSTANDING

How many interactions do we need to fully understand the model?



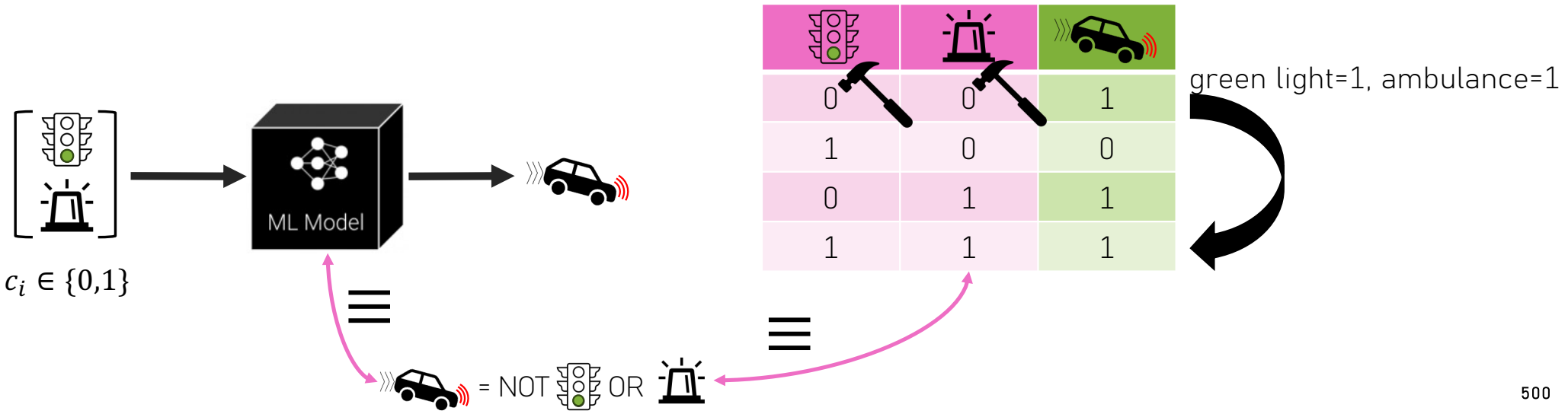
$c_i \in \{0,1\}$



THE COMPLEXITY OF UNDERSTANDING

How many interactions do we need to fully understand the model?

We can recover the full Conditional Probability / Truth Table!



THE COMPLEXITY OF UNDERSTANDING

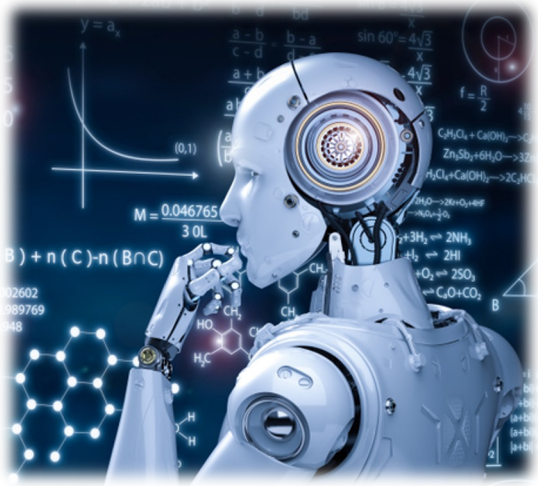
How many interactions do we need to fully understand the model?

We can recover the full Conditional Probability / Truth Table!

#interactions required to extract full CPT/TT
is exponential in #inputs!

Can we do better?

(INFORMAL) OBJECTIVE



Build a general-purpose neural model that is:

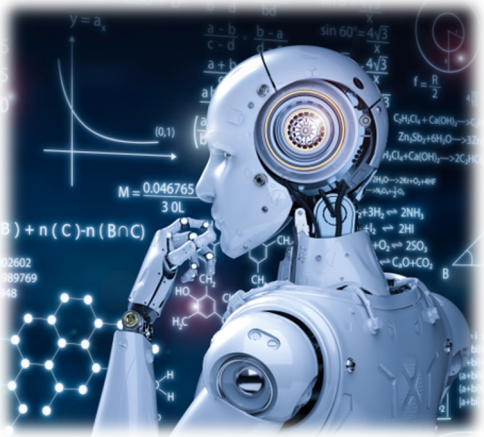
- ❑ Expressive (as DNNs)

and whose **inference mechanism** is

- ❑ Functionally transparent (we fully understand CPT/TT)
- ❑ Tractable (CPT/TT size << exp)
- ❑ Semantically transparent (concept-based)
- ❑ Causally transparent (based on non-trivial cause-effect chains)

[1] Barbiero, Pietro, et al. "Neural Interpretable Reasoning." arXiv preprint (2025).

PLAN



- General methods (foundations)
- Neural Interpretable Reasoning paradigm
- Grounding NIR (Concept Memory Reasoning example)
- Causal reasoning (metrics and models)

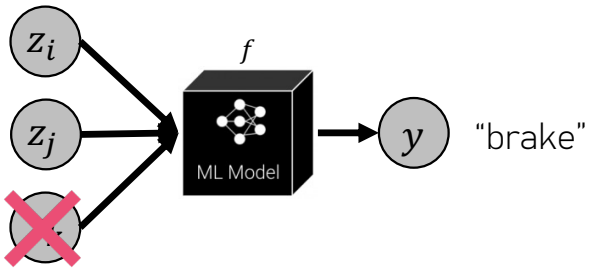
METHODS

Functional transparency: understanding of inference mechanism is tractable

$$|\mathrm{CPT}(f)| \ll \exp(|Z|)$$

Methods:

- Filter irrelevant features out



[1] Barbiero, Pietro, et al. "Neural Interpretable Reasoning." arXiv preprint (2025).

METHODS

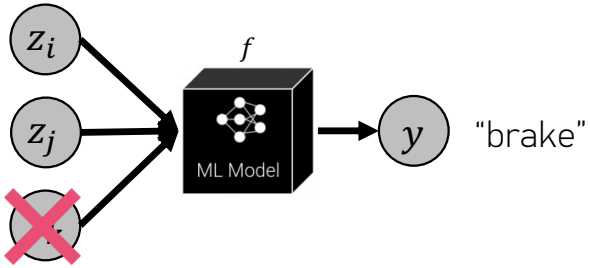
Functional transparency: understanding of inference mechanism is tractable

$$|\mathrm{CPT}(f)| \ll \exp(|Z|)$$

Methods:

- Filter irrelevant features out
- Re-parametrize global CPT as a mixture of simple (e.g., linear) "local CPTs"

SENN LENs



Global CPT

$$\text{NOT } Z_i \text{ OR } Z_j = \text{OR} \left\{ \begin{array}{c} \text{NOT } Z_i \\ Z_j \end{array} \right\}$$

Local CPT

[1] Barbiero, Pietro, et al. "Neural Interpretable Reasoning." arXiv preprint (2025).

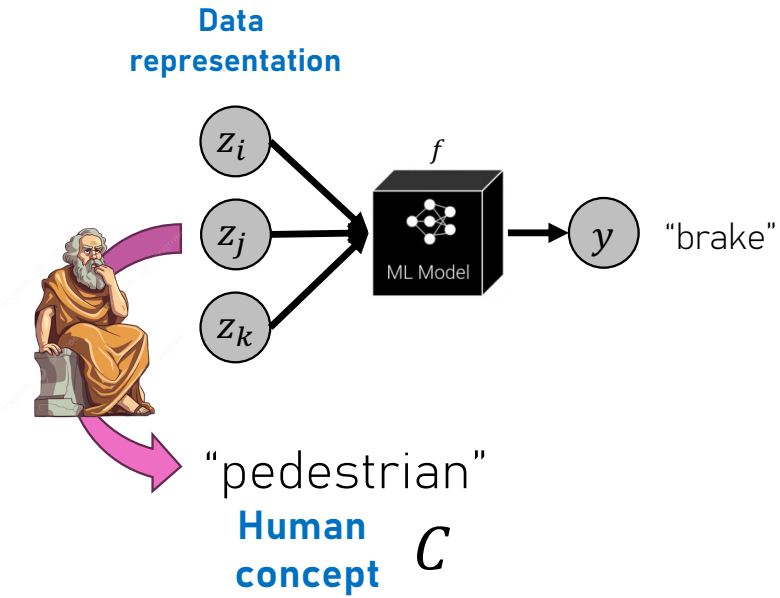
METHODS

Semantic transparency: input features are aligned with human semantics

Methods:

- Work with tabular data
- Use concept-based approach

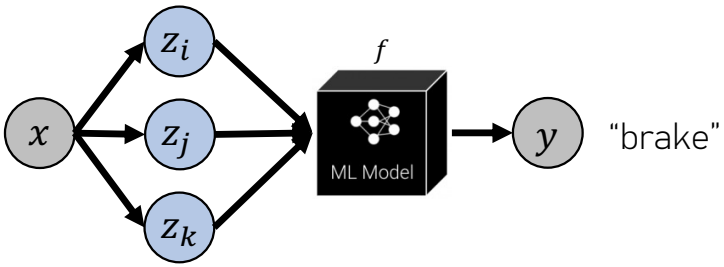
… and then write the model as $P(Y \mid C)$!



[1] Barbiero, Pietro, et al. "Neural Interpretable Reasoning." arXiv preprint (2025).

METHODS

General-purpose: model design should be applicable to any data type



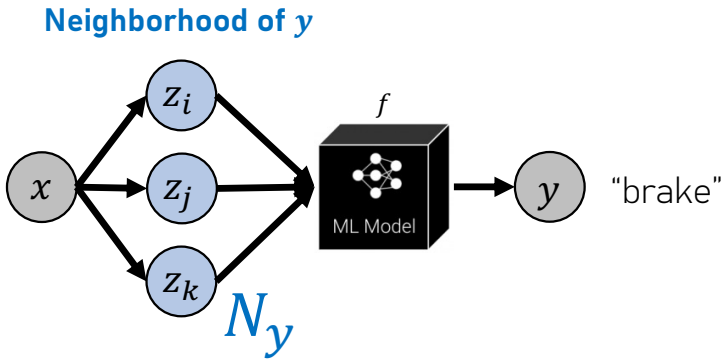
[1] Barbiero, Pietro, et al. "Neural Interpretable Reasoning." arXiv preprint (2025).

METHODS

General-purpose: model design should be applicable to any data type

Interpretability is a Markovian property:

$$(\mathrm{Y} \perp X) \mid N_y$$



Methods:

- Re-parametrize the inputs of an ML model f without affecting its interpretability (idea behind CBMs)!

$$P(Y \mid X) = \sum_C P(Y \mid C) P(C \mid X)$$

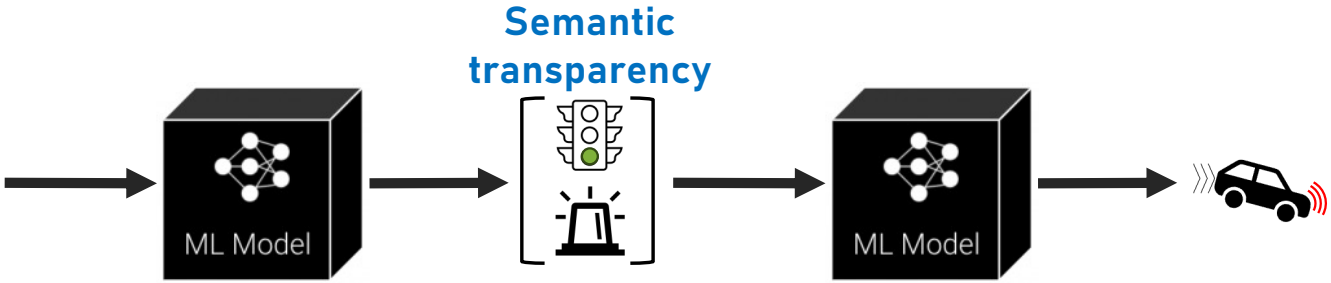
Introducing x does not affect the understanding of f!

[1] Barbiero, Pietro, et al. "Neural Interpretable Reasoning." arXiv preprint (2025).

SEMANTIC & FUNCTIONAL OPACITY

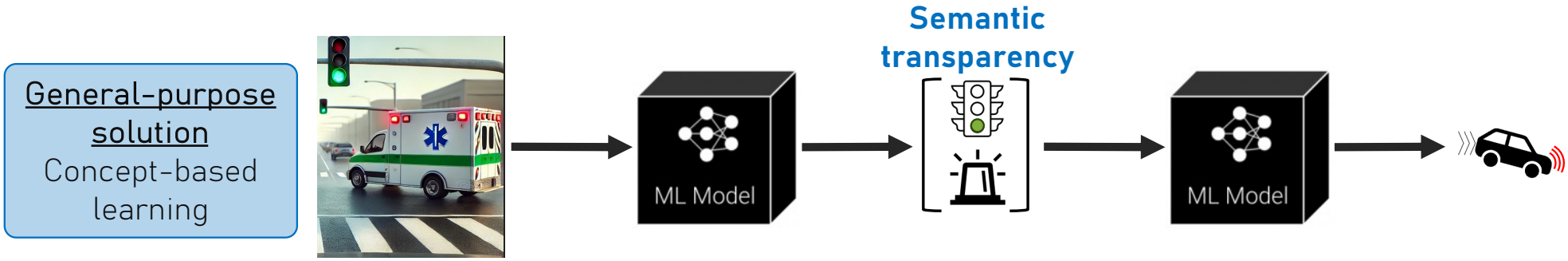
Semantic opacity: data representations are not aligned with human concepts

General-purpose solution
Concept-based learning

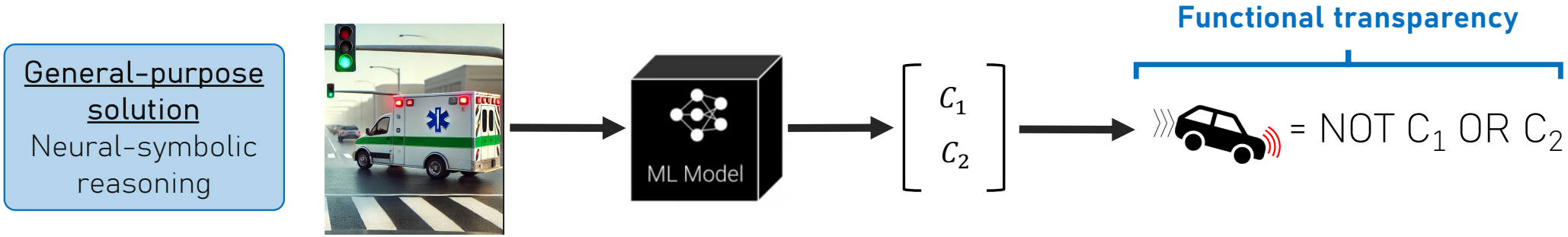


SEMANTIC & FUNCTIONAL OPACITY

Semantic opacity: data representations are not aligned with human concepts



Functional opacity: CPT is unknown or intractable to reconstruct



SEMANTIC & FUNCTIONAL OPACITY

Does semantic transparency imply functional transparency?

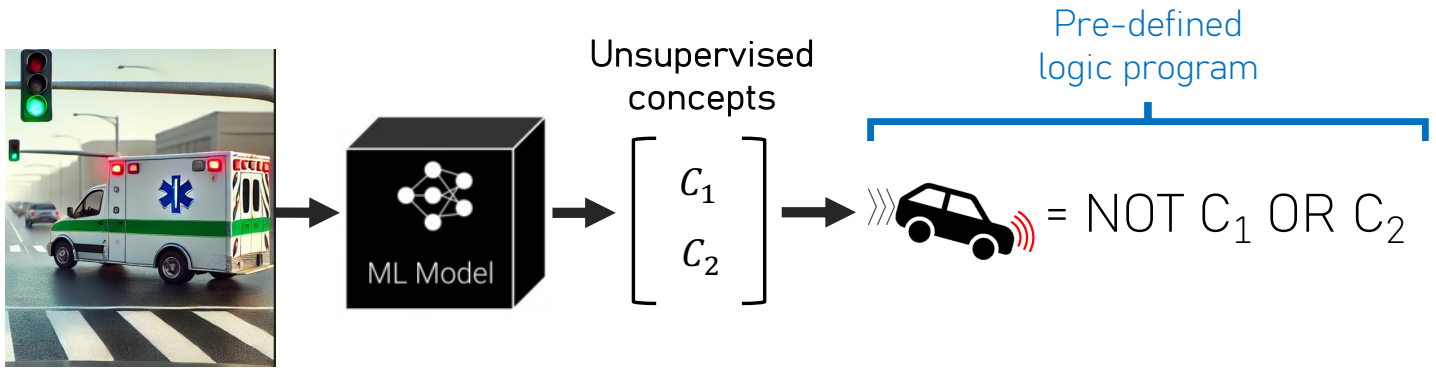
In theory it does, but…

… with n inputs the size of a TT is 2^n!
→ intractable!

SEMANTIC & FUNCTIONAL OPACITY

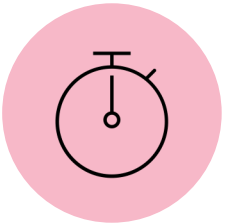
Does functional transparency imply semantic transparency?

Potential solution: DeepProbLog

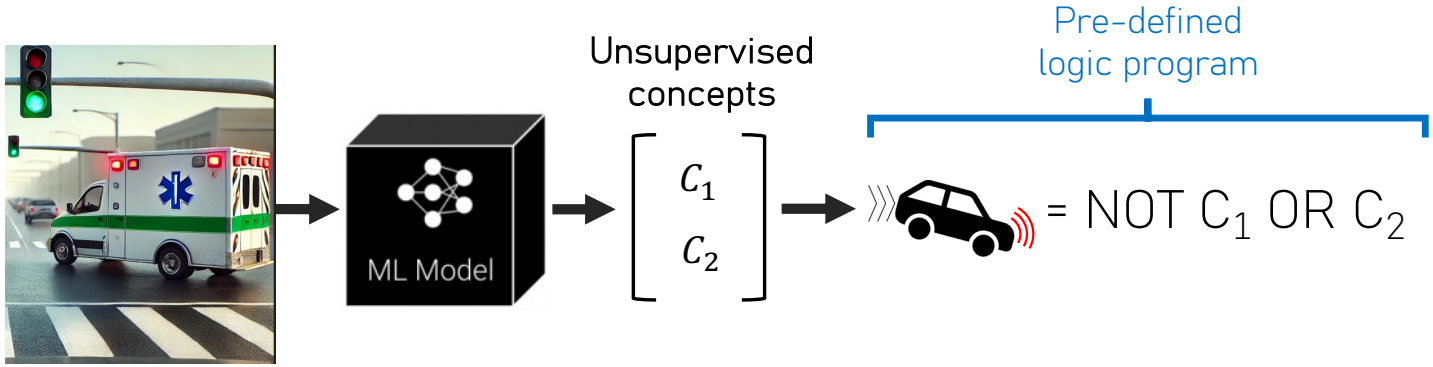


[1] Manhaeve, Robin, et al. "Deepproblog: Neural probabilistic logic programming." *NeurIPS* 2018.

FUNCTIONAL ⇒ SEMANTIC TRANSP.?



Can we prove that the classification head is safe?



If yes, raise your hand!

Disclaimer: Unfortunately for you, I'm always right during the lecture—but I'm happy to be proven wrong afterward! ☺

FUNCTIONAL ⇒ SEMANTIC TRANSP.?

Can we prove that the classification head is safe?

ML Model

Unsupervised concepts

$$\begin{bmatrix} C_1 \\ C_2 \end{bmatrix}$$

Pre-defined logic program

= NOT C_1 OR C_2

Safe!

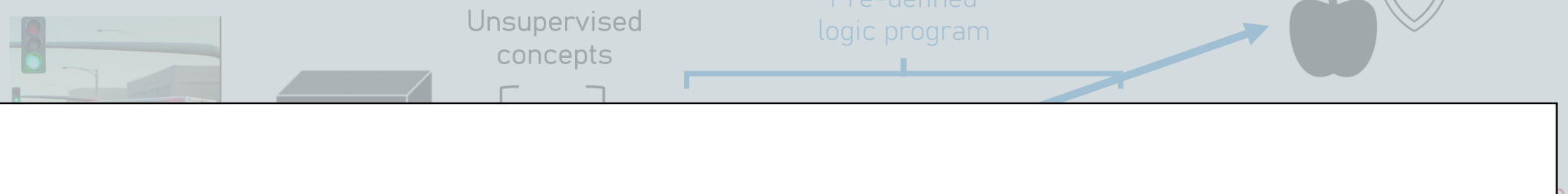
Danger!

Safety depends on concepts' meaning!

Disclaimer: Unfortunately for you, I'm always right during the lecture—but I'm happy to be proven wrong afterward! ☺

FUNCTIONAL ⇒ SEMANTIC TRANSP.?

Can we prove that the classification head is safe?



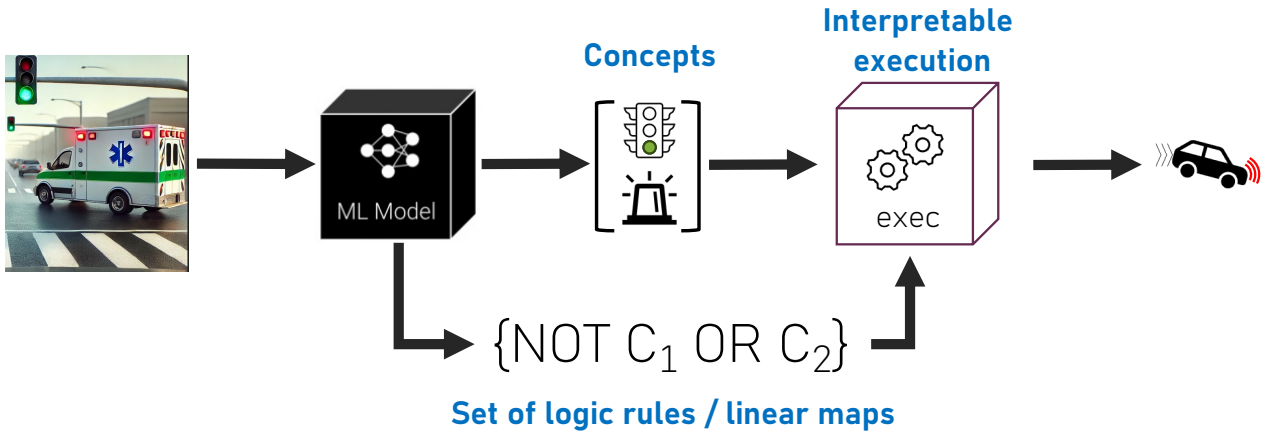
Can we combine concept-based learning with symbolic reasoning?

Disclaimer: Unfortunately for you, I'm always right during the lecture—but I'm happy to be proven wrong afterward! ☺

NEURAL INTERPRETABLE REASONING

Proposed Solution

Step 1: DNN generates both concept activations & rule parameters *(neural generation)*
Step 2: Symbolic engine executes the rule using concept activations *(interpretable execution)*



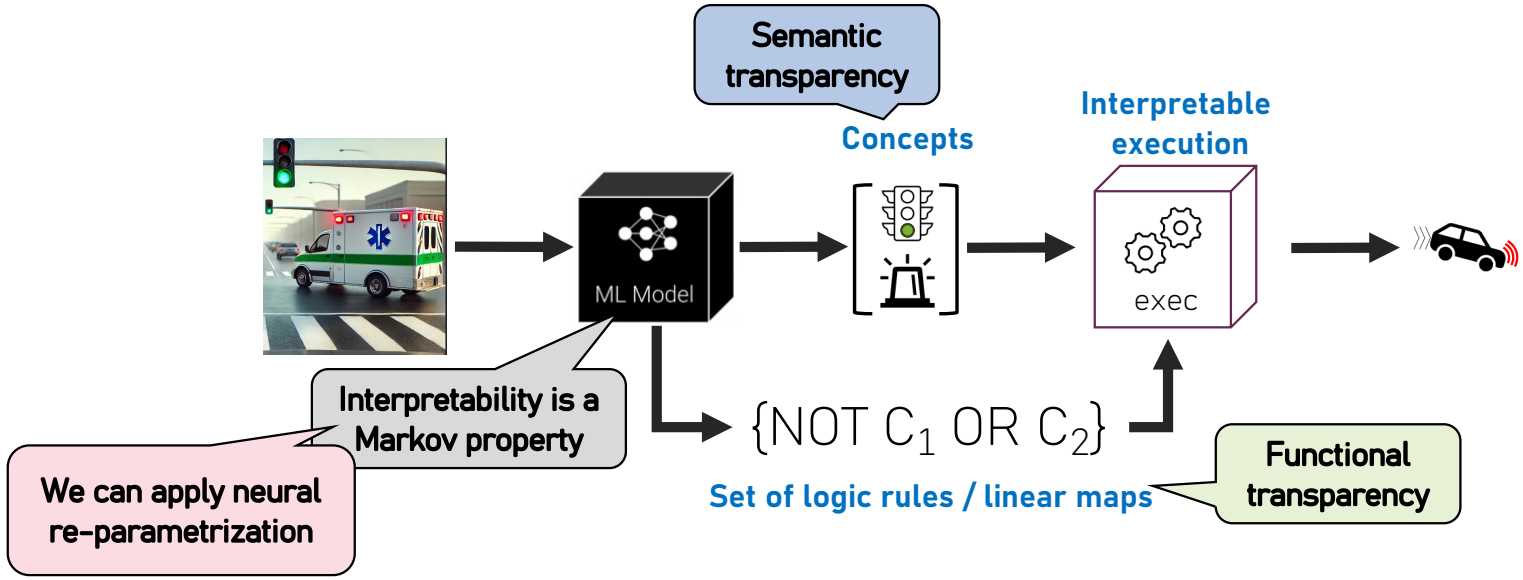
ICML23 NeurIPS24

[1] Barbiero, Pietro, et al. "Interpretable neural-symbolic concept reasoning." International Conference on Machine Learning. PMLR, 2023.
[2] Debot, David, et al. "Interpretable concept-based memory reasoning." NeurIPS 2024.

NEURAL INTERPRETABLE REASONING

Proposed Solution

$$P(Y \mid X) = P(Y \mid C; \vee \{W_i\})P(C, \{W_i\} \mid X)$$



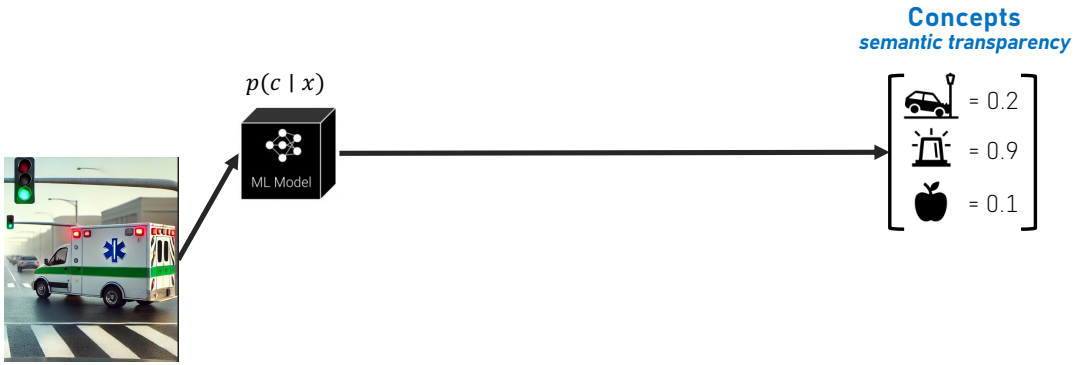
ICML23 NeurIPS24

[1] Barbiero, Pietro, et al. "Interpretable neural-symbolic concept reasoning." International Conference on Machine Learning. PMLR, 2023.
[2] Debot, David, et al. "Interpretable concept-based memory reasoning." NeurIPS 2024.

CONCEPT-BASED MEMORY REASONING

Proposed Solution

Step 1: DNN predicts concept activations

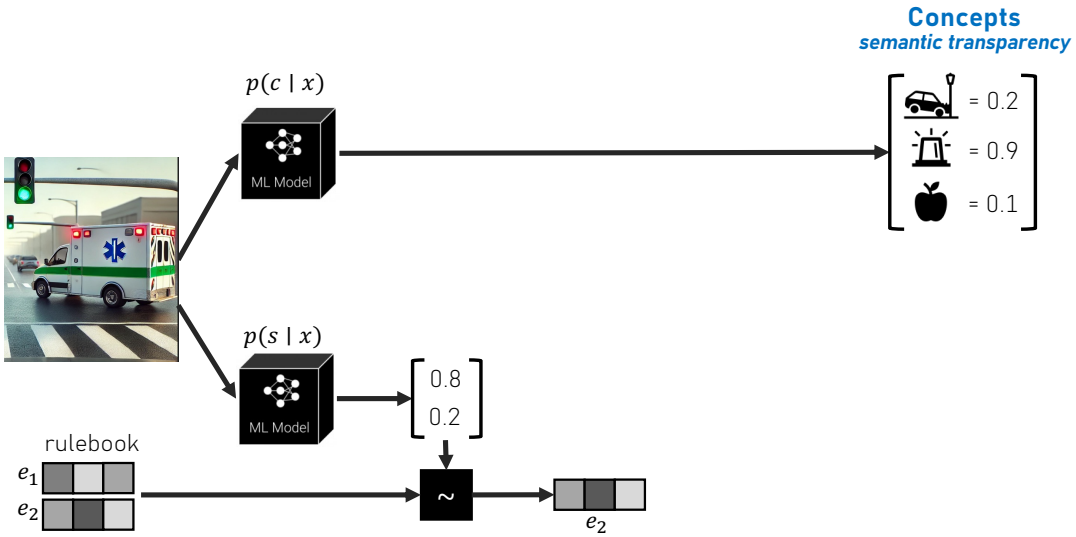


[1] Debot, David, et al. "Interpretable concept-based memory reasoning." NeurIPS 2024.

CONCEPT-BASED MEMORY REASONING

Proposed Solution

Step 2: DNN predicts embedding to be selected from the latent rulebook

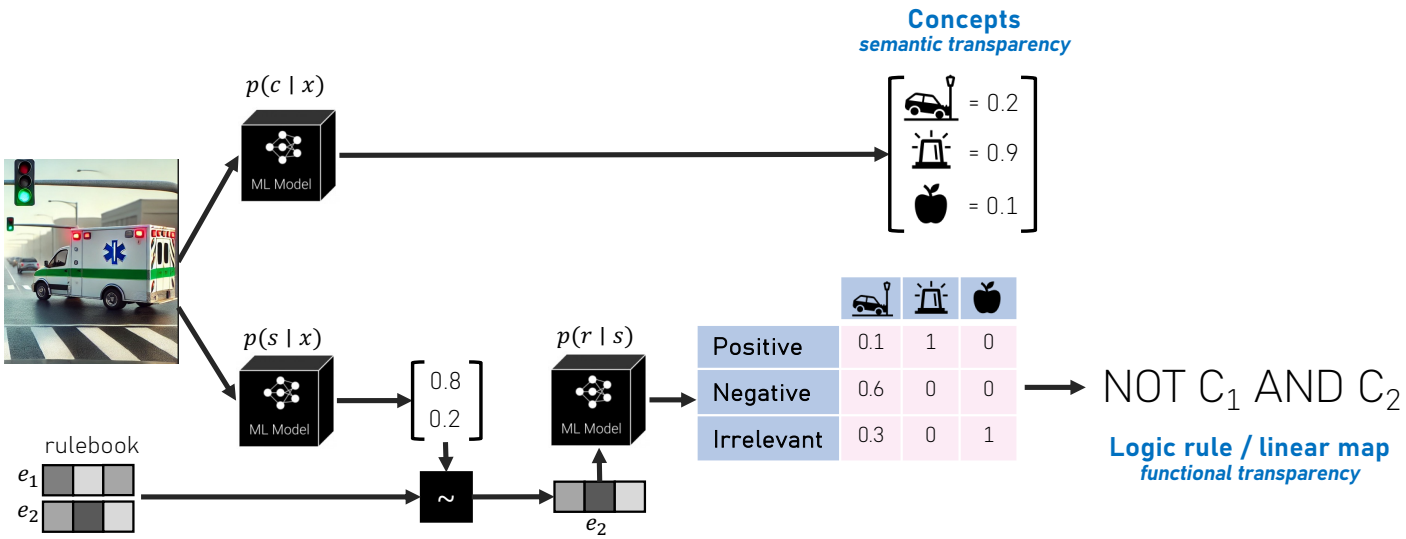


[1] Debot, David, et al. "Interpretable concept-based memory reasoning." NeurIPS 2024.

CONCEPT-BASED MEMORY REASONING

Proposed Solution

Step 3: DNN decodes selected embedding into 3 states: positive, negative, irrelevant



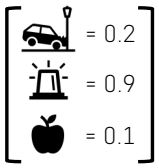
[1] Debot, David, et al. "Interpretable concept-based memory reasoning." NeurIPS 2024.

CONCEPT-BASED MEMORY REASONING

Proposed Solution (details)

Step 4: Execute rule using rule states and concept activations

Concepts



Rule states

NOT C_1 AND C_2

[1] Debot, David, et al. "Interpretable concept-based memory reasoning." NeurIPS 2024.

CONCEPT-BASED MEMORY REASONING

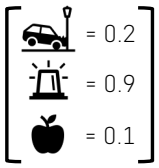
Proposed Solution (details)

Step 4: Execute rule using rule states and concept activations

Execution of rule states on concept activations

$$\mathbb{I}[r_i = \text{positive}]p(c_i \mid x) + \mathbb{I}[r_i = \text{negative}]p(\neg c_i \mid x) + \mathbb{I}[r_i = \text{irrelevant}]$$

Concepts



Rule states

NOT C_1 AND C_2

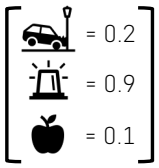
[1] Debot, David, et al. "Interpretable concept-based memory reasoning." NeurIPS 2024.

CONCEPT-BASED MEMORY REASONING

Proposed Solution (details)

Step 4: Execute rule using rule states and concept activations

Concepts



Rule states

NOT C_1 AND C_2

Execution of rule states on concept activations

$$\mathbb{I}[r_i = \text{positive}]p(c_i \mid x) + \mathbb{I}[r_i = \text{negative}]p(\neg c_i \mid x) + \mathbb{I}[r_i = \text{irrelevant}]$$

$$0 \times p(c_i \mid x) + 1 \times p(\neg c_i \mid x) + 0 = p(\neg c_i \mid x) = (1 - 0.2) = 0.8$$

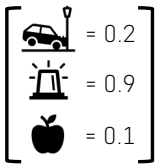
[1] Debot, David, et al. "Interpretable concept-based memory reasoning." NeurIPS 2024.

CONCEPT-BASED MEMORY REASONING

Proposed Solution (details)

Step 4: Execute rule using rule states and concept activations

Concepts



Rule states

NOT C_1 AND C_2

Execution of rule states on concept activations

$$\mathbb{I}[r_i = \text{positive}]p(c_i \mid x) + \mathbb{I}[r_i = \text{negative}]p(\neg c_i \mid x) + \mathbb{I}[r_i = \text{irrelevant}]$$

$$0 \times p(c_i \mid x) + 1 \times p(\neg c_i \mid x) + 0 = p(\neg c_i \mid x) = (1 - 0.2) = 0.8$$

$$1 \times p(c_i \mid x) + 0 \times p(\neg c_i \mid x) + 0 = p(c_i \mid x) = 0.9$$

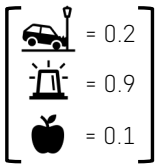
[1] Debot, David, et al. "Interpretable concept-based memory reasoning." NeurIPS 2024.

CONCEPT-BASED MEMORY REASONING

Proposed Solution (details)

Step 4: Execute rule using rule states and concept activations

Concepts



Rule states

NOT C_1 AND C_2

Execution of rule states on concept activations

$$\mathbb{I}[r_i = \text{positive}]p(c_i \mid x) + \mathbb{I}[r_i = \text{negative}]p(\neg c_i \mid x) + \mathbb{I}[r_i = \text{irrelevant}]$$

$$0 \times p(c_i \mid x) + 1 \times p(\neg c_i \mid x) + 0 = p(\neg c_i \mid x) = (1 - 0.2) = 0.8$$

$$1 \times p(c_i \mid x) + 0 \times p(\neg c_i \mid x) + 0 = p(c_i \mid x) = 0.9$$

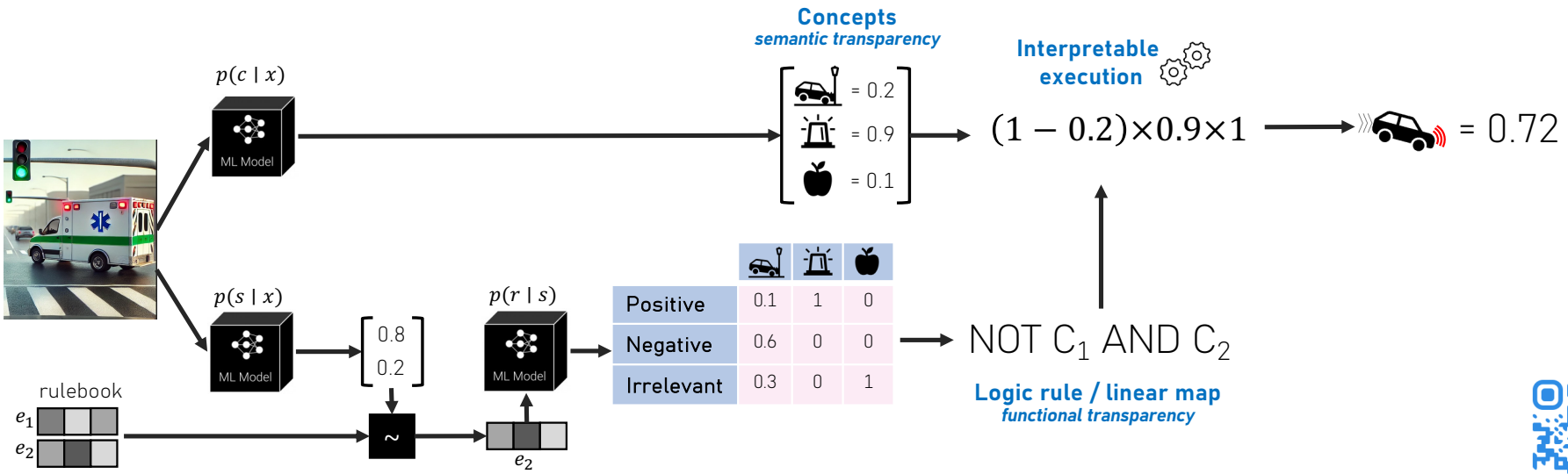
$$0 \times p(c_i \mid x) + 0 \times p(\neg c_i \mid x) + 1 = 1$$

[1] Debot, David, et al. "Interpretable concept-based memory reasoning." NeurIPS 2024.

CONCEPT-BASED MEMORY REASONING

Proposed Solution

Step 5: Execute the rule combining concept states and activations to predict the output label

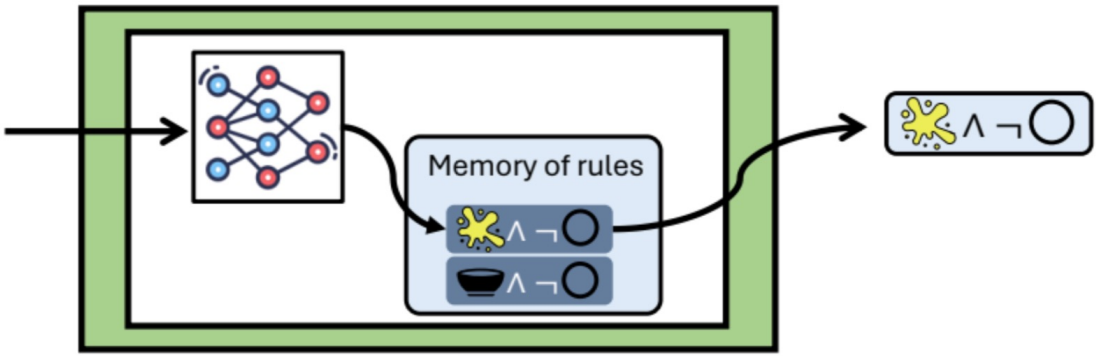


[1] Debot, David, et al. "Interpretable concept-based memory reasoning." NeurIPS 2024.

CONCEPT-BASED MEMORY REASONING

CMR has 3 key features:

- Universal approximator akin to opaque DNNs (Theorem 4.1)



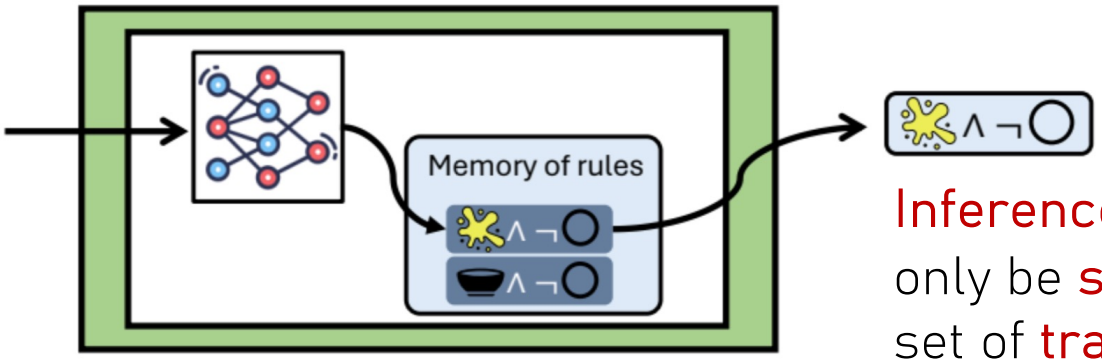
Mixture of linear models where the mixing coeff. are parametrized by the selector

[1] Debot, David, et al. "Interpretable concept-based memory reasoning." NeurIPS 2024.

CONCEPT-BASED MEMORY REASONING

CMR has 3 key features:

- Universal approximator akin to opaque DNNs (Theorem 4.1)
- Minimize #relevant concepts → understanding is tractable



Inference mechanisms can only be selected from a finite set of transparent rules!

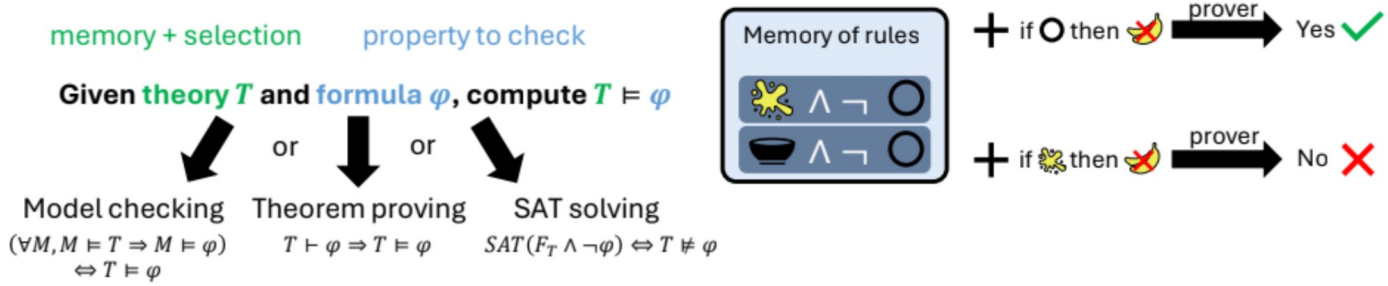
[1] Debot, David, et al. "Interpretable concept-based memory reasoning." NeurIPS 2024.

CONCEPT-BASED MEMORY REASONING

CMR has 3 key features:

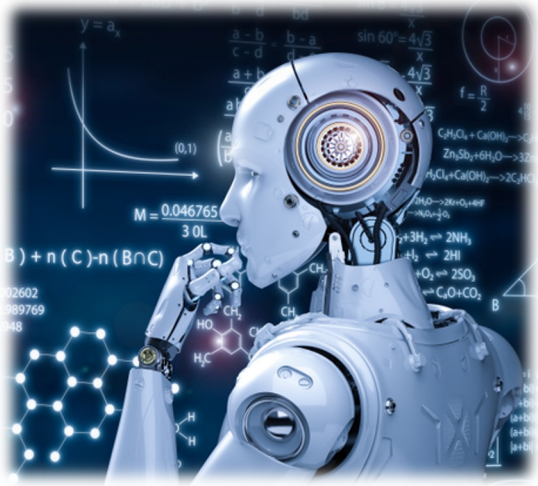
- Universal approximator akin to opaque DNNs (Theorem 4.1)
- Minimize #relevant concepts → understanding is tractable
- The concept memory allows formal verification of properties

"Does a property hold no matter which rule is selected?"



[1] Debot, David, et al. "Interpretable concept-based memory reasoning." NeurIPS 2024.

(INFORMAL) OBJECTIVE



Build a **general-purpose** neural model that is:

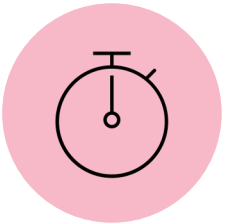
☑ Expressive (as DNNs)

and whose **inference mechanism** is

☑ Functionally transparent (we fully understand CPT/TT)
☑ Tractable (CPT/TT size << exp)
☑ Semantically transparent (concept-based)
❑ Causally transparent (based on non-trivial cause-effect chains)

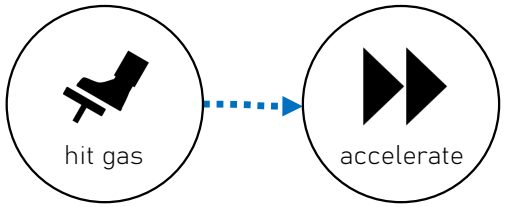
[1] Barbiero, Pietro, et al. "Neural Interpretable Reasoning." arXiv preprint (2025).

COMPARING BOTTLENECKS

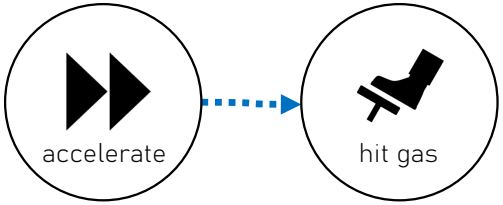


Is CBM #1 more causally transparent than CBM #2?

CBM #1



CBM #2



If yes, raise your hand!

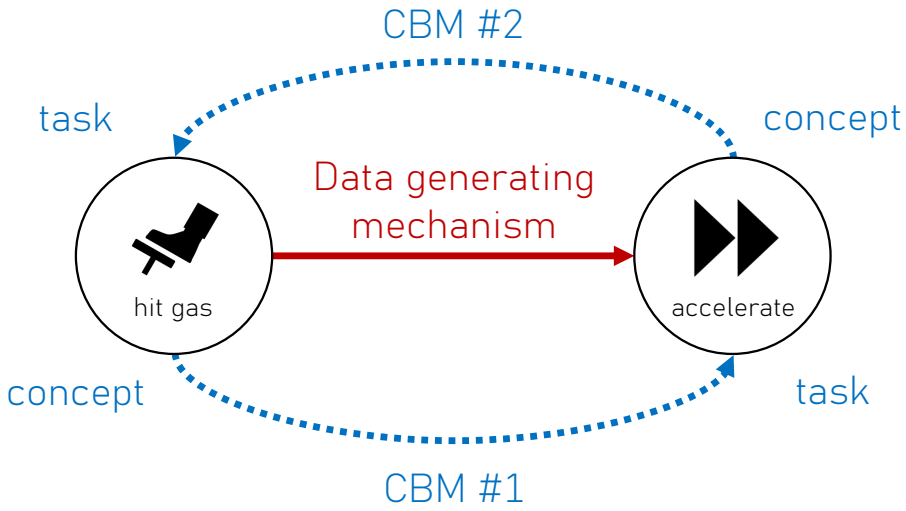
Disclaimer: Unfortunately for you, I'm always right during the lecture—but I'm happy to be proven wrong afterward! ☺

CAUSAL OPACITY

- Causal reliability: discover causal mechanisms of the data generating process

CAUSAL OPACITY

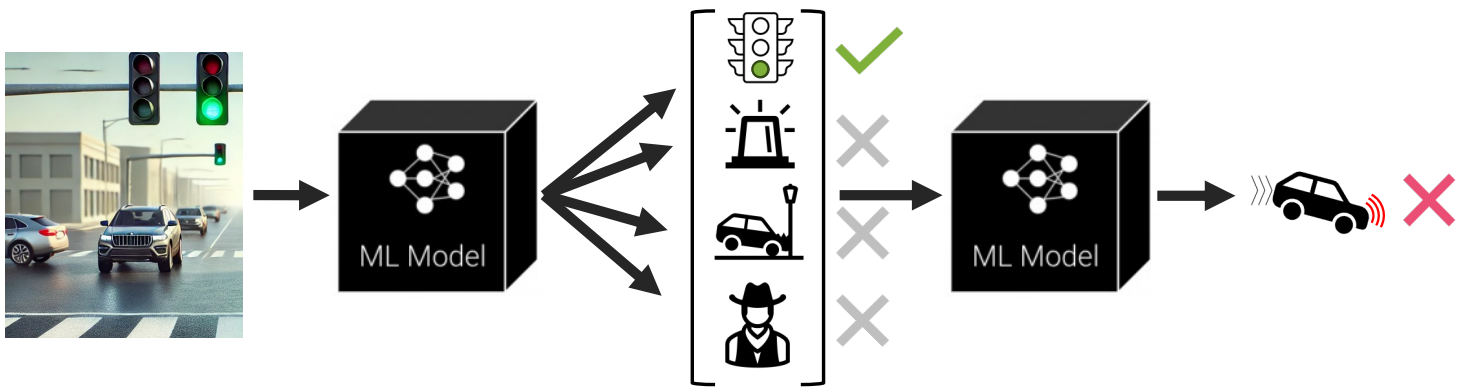
- Causal reliability: discover causal mechanisms of the data generating process
- Causal opacity: discover causal mechanism of a model's inference process



CBM #1 is potentially equally transparent w.r.t. CBM #2!

MEASURING CAUSAL EFFECTS

CBMs can answer association queries (duh…)



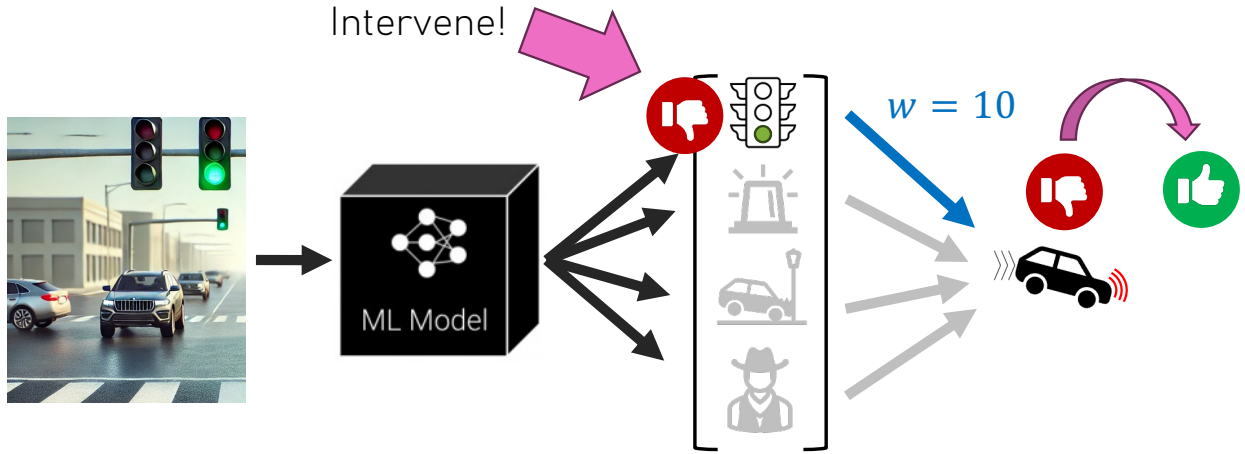
Association

What if the model sees a green light?

$P(brake \mid light)$

MEASURING CAUSAL EFFECTS

Sometimes intervening on wrongly predicted concepts helps...

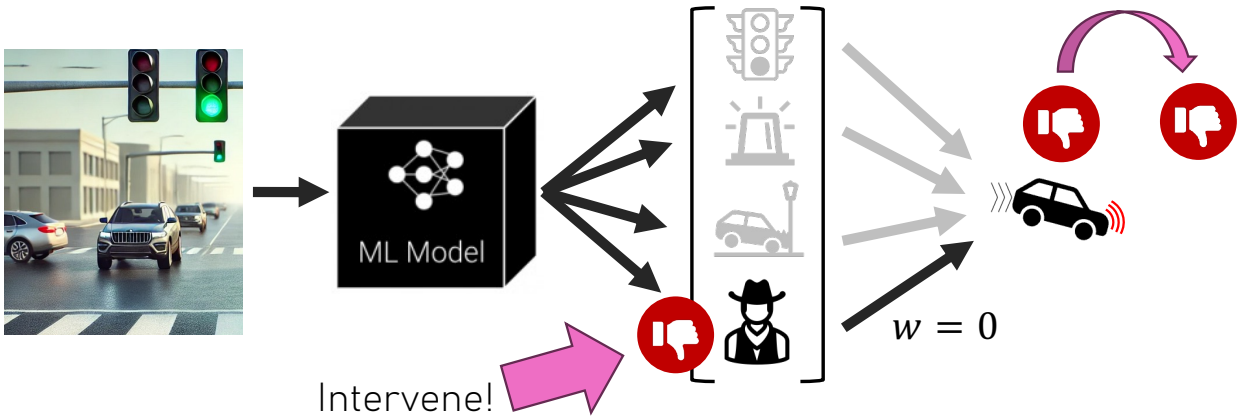


MEASURING CAUSAL EFFECTS

Sometimes intervening on wrongly predicted concepts helps…

and sometimes it doesn't! 😢

Causal analysis can provide us with insights!



Yes, but…

MEASURING CAUSAL EFFECTS

Sometimes intervening on wrongly predicted concepts helps...

and sometimes it doesn't! 😢

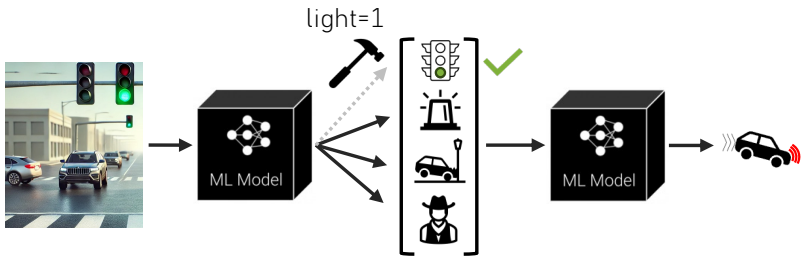
Can we measure the causal influence of a concept on the task?

Yes, but...

CAUSAL CONCEPT EFFECT

Proposed Solution

Step 1: Compute expected value of the task with $do(c_i = 1)$



$$\mathbb{E}[\,\text{brake} \mid do(\text{light} = 1)\,] = 0.2$$

Intervention

What if I set the light color to red?

$P(brake \mid do(light))$

Association

What if the model sees a green light?

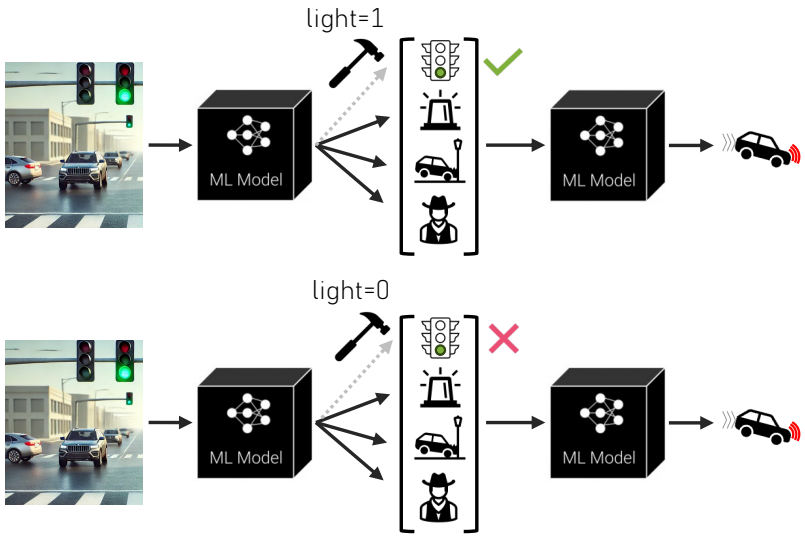
$P(brake \mid light)$

[1] Goyal, Yash, et al. "Explaining classifiers with causal concept effect (cace)." arXiv preprint 2019.

CAUSAL CONCEPT EFFECT

Proposed Solution

Step 2: Compute expected value of the task with $do(c_i = 0)$



$$\mathbb{E}[\,\text{brake} \mid do(\text{light} = 1)\,] = 0.2$$

$$\mathbb{E}[\,\text{brake} \mid do(\text{light} = 0)\,] = 1$$

Intervention

What if I set the light color to red?

$P(brake \mid do(light))$

Association

What if the model sees a green light?

$P(brake \mid light)$

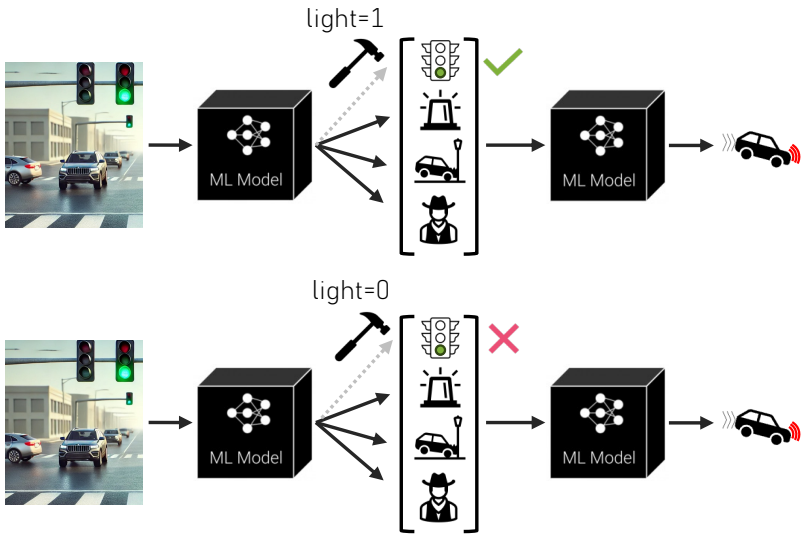
[1] Goyal, Yash, et al. "Explaining classifiers with causal concept effect (cace)." arXiv preprint 2019.

CAUSAL CONCEPT EFFECT

Proposed Solution

Step 3: Compute difference of expected values: absolute value is proportional to causal effect

$$\text{CaCE} = \mathbb{E}[\,\text{brake} \mid do(\text{light} = 1)\,] - \mathbb{E}[\,\text{brake} \mid do(\text{light} = 0)\,] = -0.8$$



$$\mathbb{E}[\,\text{brake} \mid do(\text{light} = 1)\,] = 0.2$$

$$\mathbb{E}[\,\text{brake} \mid do(\text{light} = 0)\,] = 1$$

Intervention

What if I set the light color to red?

$P(brake \mid do(light))$

Association

What if the model sees a green light?

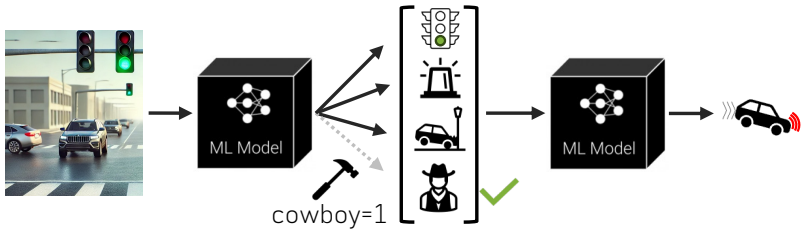
$P(brake \mid light)$

[1] Goyal, Yash, et al. "Explaining classifiers with causal concept effect (cace)." arXiv preprint 2019.

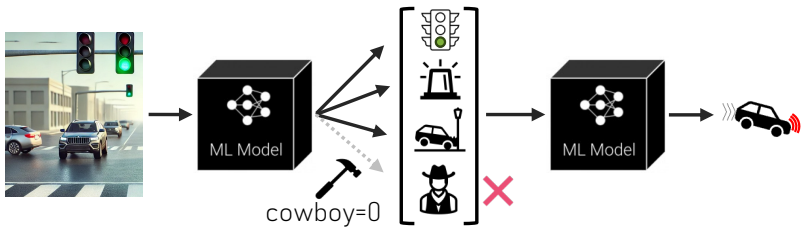
CAUSAL CONCEPT EFFECT

Proposed Solution

Step 3: Compute difference of expected values: absolute value is proportional to causal effect



$$\mathbb{E}[\,\text{brake} \mid do(\text{cowboy} = 1)\,] = 0.5$$



$$\mathbb{E}[\,\text{brake} \mid do(\text{cowboy} = 0)\,] = 0.5$$

Intervention

What if I set the light color to red?

$P(brake \mid do(light))$

Association

What if the model sees a green light?

$P(brake \mid light)$

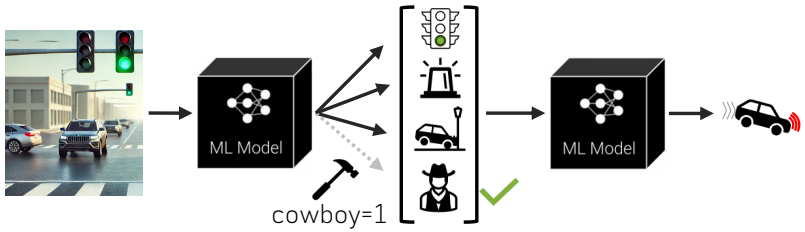
[1] Goyal, Yash, et al. "Explaining classifiers with causal concept effect (cace)." arXiv preprint 2019.

CAUSAL CONCEPT EFFECT

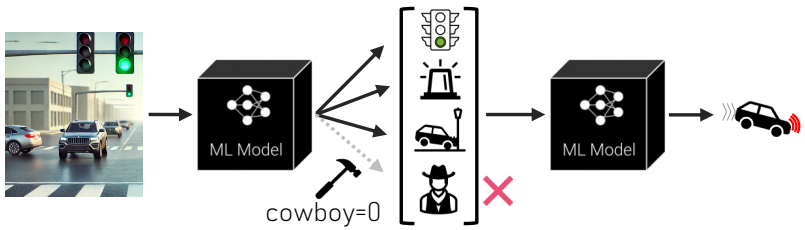
Proposed Solution

Step 3: Compute difference of expected values: absolute value is proportional to causal effect

$$\text{CaCE} = \mathbb{E}[\,\text{brake} \mid do(\text{cowboy} = 1)\,] - \mathbb{E}[\,\text{brake} \mid do(\text{cowboy} = 0)\,] = 0$$



$$\mathbb{E}[\,\text{brake} \mid do(\text{cowboy} = 1)\,] = 0.5$$



$$\mathbb{E}[\,\text{brake} \mid do(\text{cowboy} = 0)\,] = 0.5$$

Intervention

What if I set the light color to red?

$P(brake \mid do(light))$

Association

What if the model sees a green light?

$P(brake \mid light)$

[1] Goyal, Yash, et al. "Explaining classifiers with causal concept effect (cace)." arXiv preprint 2019.

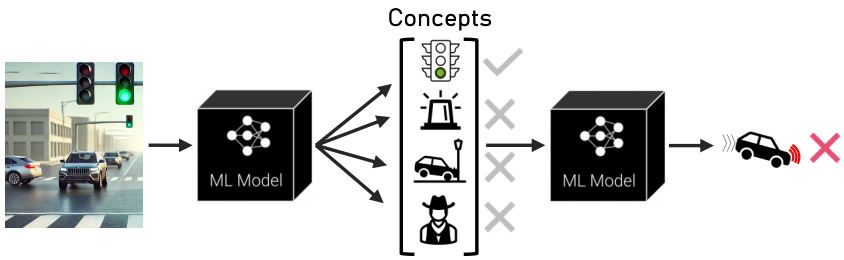
COUNTERFACTUAL CBMS

ICLR25

ICML22

Limitation Being Addressed

CBMs cannot answer counterfactual queries!



Counterfactual

What would have been predicted in the same circumstance had a car crash be seen?

$$P(brake \mid light, crash)$$

Intervention

What if I set the light color to red?

$$P(brake \mid do(light))$$

Association

What if the model sees a green light?

$$P(brake \mid light)$$

[1] Dominici, Gabriele, et al. "Counterfactual Concept Bottleneck Models." ICLR 2025.

[2] Abid, Abubakar, et al. "Meaningfully debugging model mistakes using conceptual counterfactual explanations." International Conference on Machine Learning. PMLR, 2022.

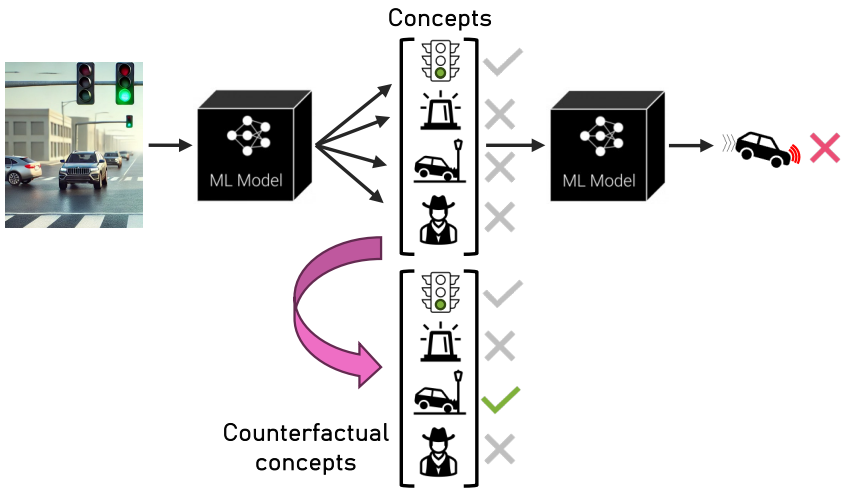
COUNTERFACTUAL CBMS

ICLR25

ICML22

Proposed Solution

Step 1: Generate counterfactual concept activations



Counterfactual

What would have been predicted in the same circumstance had a car crash be seen?

$$P(brake \mid light, crash)$$

Intervention

What if I set the light color to red?

$$P(brake \mid do(light))$$

Association

What if the model sees a green light?

$$P(brake \mid light)$$

[1] Dominici, Gabriele, et al. "Counterfactual Concept Bottleneck Models." ICLR 2025.

[2] Abid, Abubakar, et al. "Meaningfully debugging model mistakes using conceptual counterfactual explanations." International Conference on Machine Learning. PMLR, 2022.

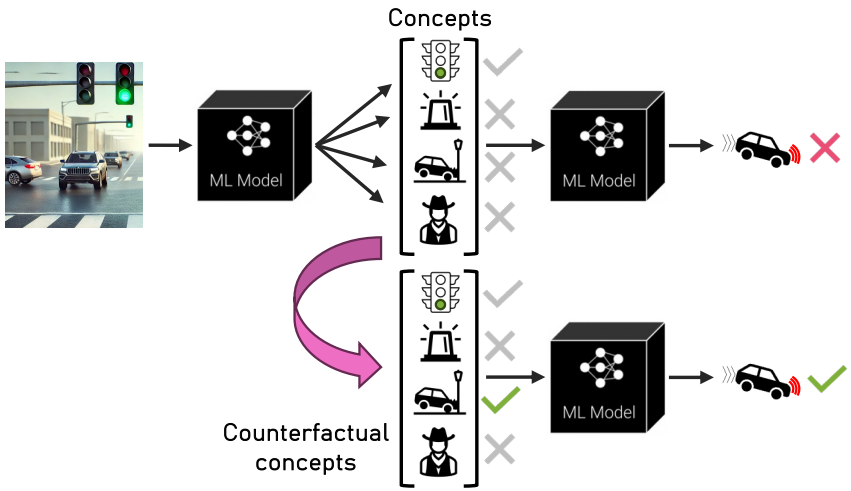
COUNTERFACTUAL CBMS

ICLR25

ICML22

Proposed Solution

Step 2: Compute causal effect on the task!



Counterfactual

What would have been predicted in the same circumstance had a car crash be seen?

$$P(brake \mid light, crash)$$

Intervention

What if I set the light color to red?

$$P(brake \mid do(light))$$

Association

What if the model sees a green light?

$$P(brake \mid light)$$

[1] Dominici, Gabriele, et al. "Counterfactual Concept Bottleneck Models." ICLR 2025.

[2] Abid, Abubakar, et al. "Meaningfully debugging model mistakes using conceptual counterfactual explanations." International Conference on Machine Learning. PMLR, 2022.

DIRECT COUNTERFACTUAL DEPENDENCE

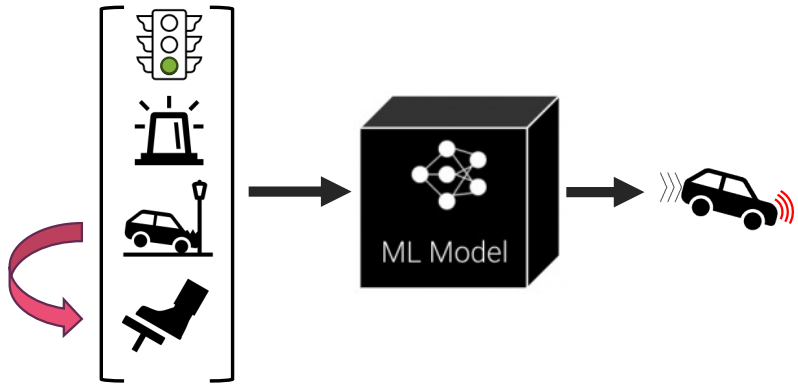
So far, we have been making 2 strong assumptions...

DIRECT COUNTERFACTUAL DEPENDENCE

So far, we have been making 2 strong assumptions:

- Concepts are mutually independent

Intervening on "car crash" does not increase the likelihood of hitting the brakes!



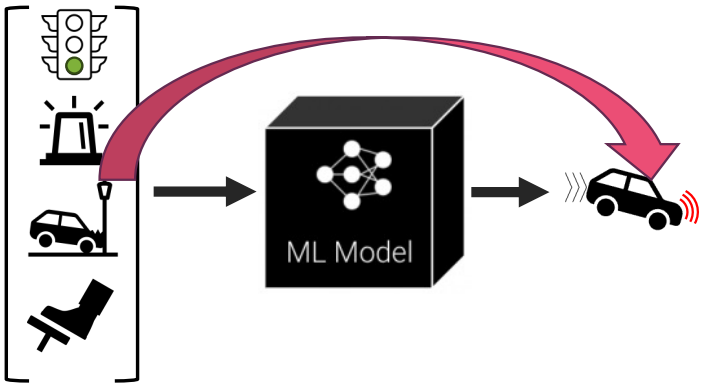
DIRECT COUNTERFACTUAL DEPENDENCE

So far, we have been making 2 strong assumptions:

- Concepts are mutually independent
- Concepts are direct causes of the task

Intervening on "car crash" does not increase the likelihood of hitting the brakes!

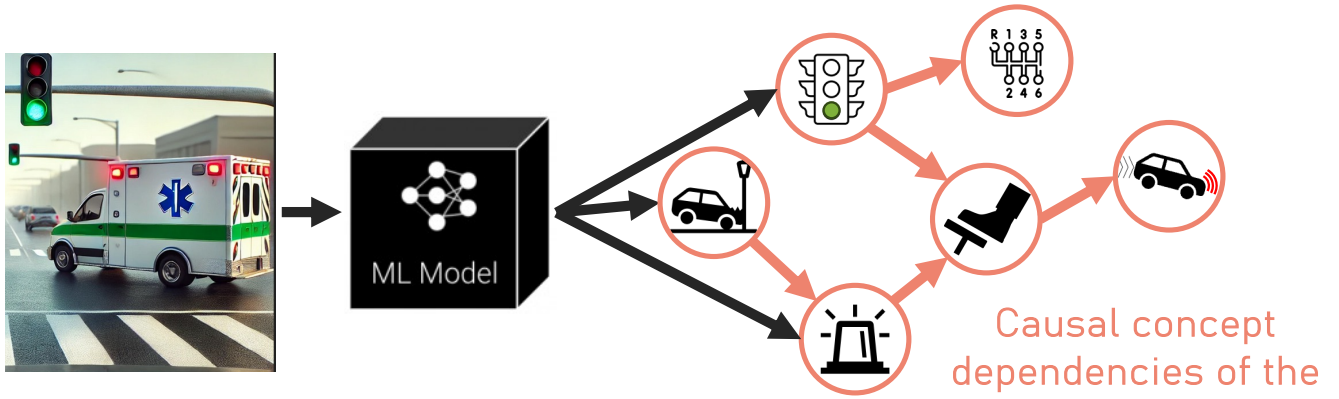
Intervening on "car crash" directly causes the car to brake!



CONCEPT GRAPH MODELS

Proposed Solution

Enforce inference through a concept graph!



ICLR25 CLeaR24

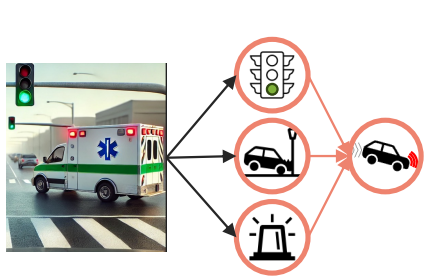
[1] Dominici, Gabriele et al. "Causal Concept Graph Models: Beyond Causal Opacity in Deep Learning." ICLR 2025.
[2] Moreira, Ricardo, et al. "Diconstruct: Causal concept-based explanations through black-box distillation." CLeaR 2024.

CONCEPT GRAPH MODELS

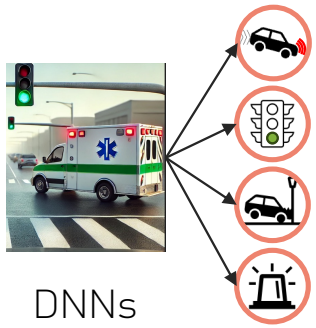
Proposed Solution

The concept graph can be:

- Given as a prior



CBMs



DNNs

ICLR25

CLeaR24

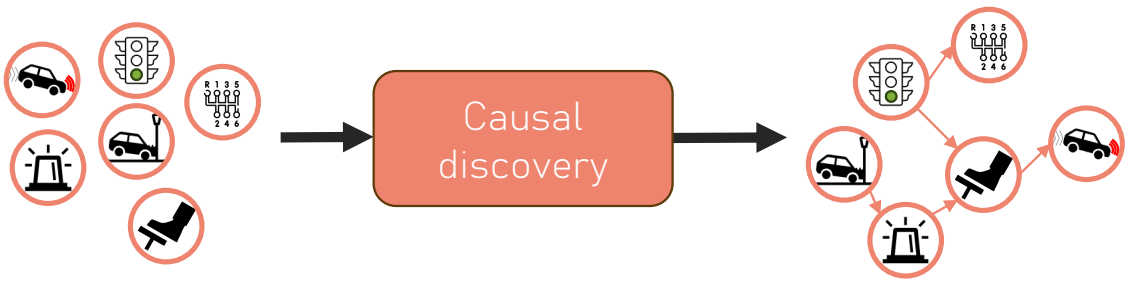
[1] Dominici, Gabriele et al. "Causal Concept Graph Models: Beyond Causal Opacity in Deep Learning." ICLR 2025.
[2] Moreira, Ricardo, et al. "Diconstruct: Causal concept-based explanations through black-box distillation." CLeaR 2024.

CONCEPT GRAPH MODELS

Proposed Solution

The concept graph can be:

- Given as a prior
- Extracted from data with causal discovery techniques



ICLR25

CLeaR24

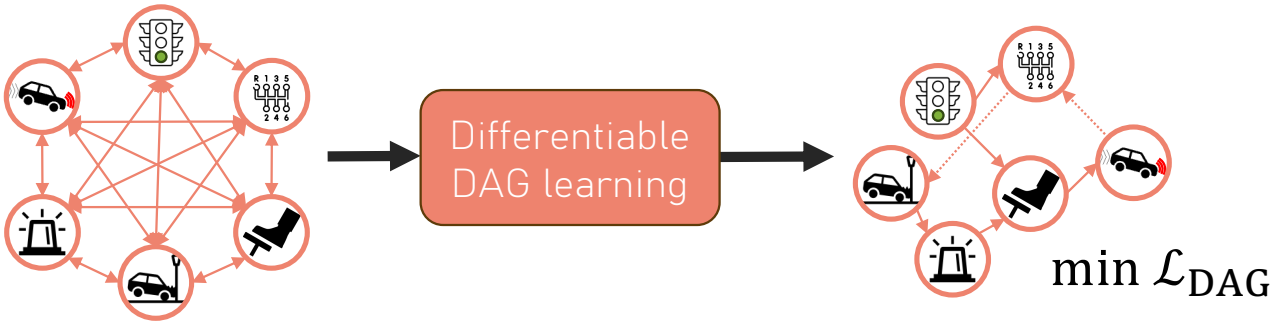
[1] Dominici, Gabriele et al. "Causal Concept Graph Models: Beyond Causal Opacity in Deep Learning." ICLR 2025.
[2] Moreira, Ricardo, et al. "Diconstruct: Causal concept-based explanations through black-box distillation." CLeaR 2024.

CONCEPT GRAPH MODELS

Proposed Solution

The concept graph can be:

- Given as a prior
- Extracted from data with causal discovery techniques
- Obtained with differentiable DAG learning



ICLR25 CLeaR24

[1] Dominici, Gabriele et al. "Causal Concept Graph Models: Beyond Causal Opacity in Deep Learning." ICLR 2025.
[2] Moreira, Ricardo, et al. "Diconstruct: Causal concept-based explanations through black-box distillation." CLeaR 2024.

NEURAL INTERPRETABLE REASONING

Build a **general-purpose** neural model that is:

☑ Expressive (as DNNs)

and whose **inference mechanism** is

☑ Semantically transparent (concept-based)
☑ Functionally transparent (we fully understand CPT/TT)
☑ Tractable (CPT/TT size << exp)
☑ Causally transparent (based on non-trivial cause-effect chains)

[1] Barbiero, Pietro, et al. "Neural Interpretable Reasoning." arXiv preprint (2025).