Everything, Everywhere, All at Once: Is Mechanistic Interpretability Identifiable?

Maxime Méloux, Silviu Maniu, François Portet, Maxime Peyrard

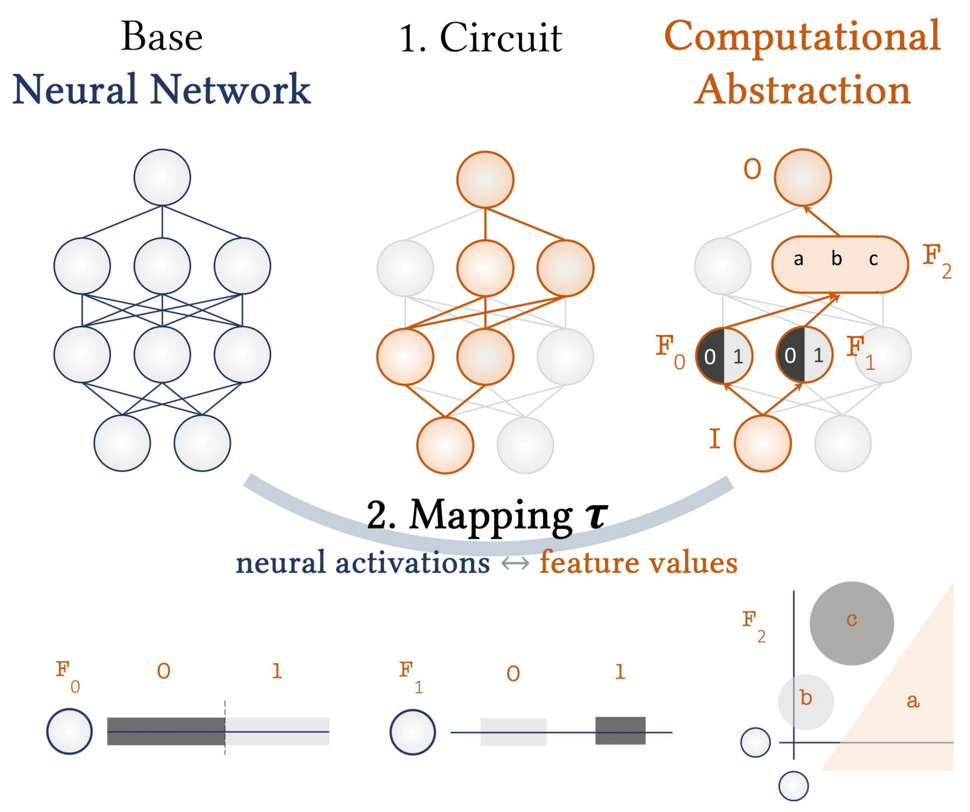
Université Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

What is Mechanistic Intepretability?

Mechanistic Interpretability (MI): reverse-engineer neural systems to uncover simple, human-interpretable algorithms embedded in the neural network structure.

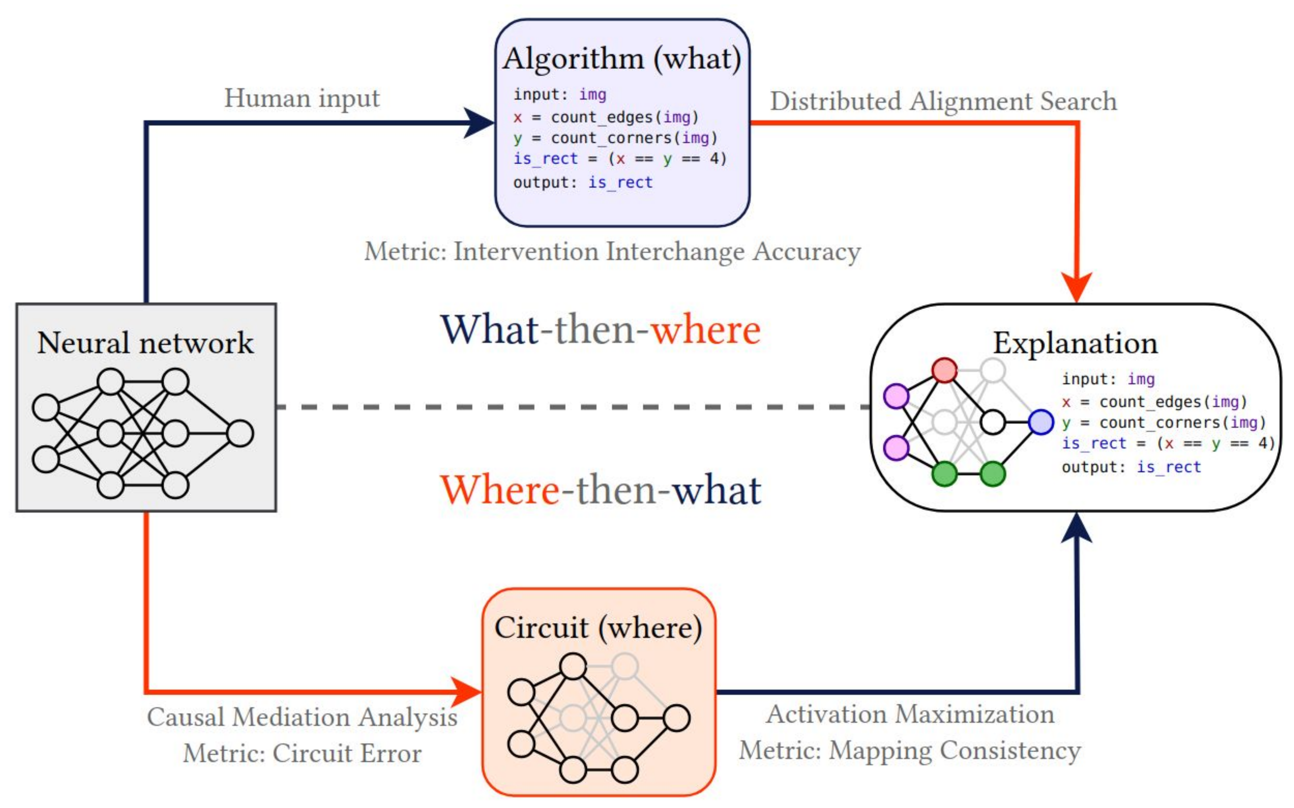
MI explanations (computational abstractions) have two parts:

- ***What*** algorithm explains a given behavior? (**mapping** of low-level activations to high-level feature values)
- ***Where*** is the algorithm embedded in the network? (**circuit**: subset of the computational graph)



MI strategies and criteria

Current techniques can be classified into two strategies: *where-then-what* and *what-then-where*.

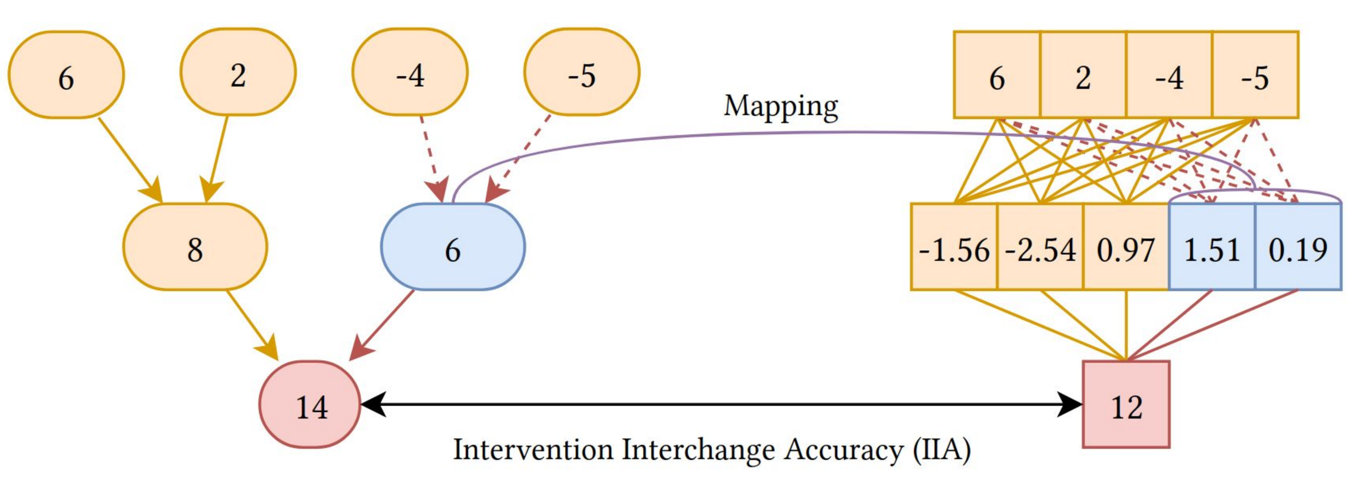


Where-then-what metrics:

- **Circuit error**: how well does the circuit replicate the behavior?
- **Mapping consistency**: does the mapping consistently align the computations in the low-level model with those in the high-level algorithm?

What-then-where metric:

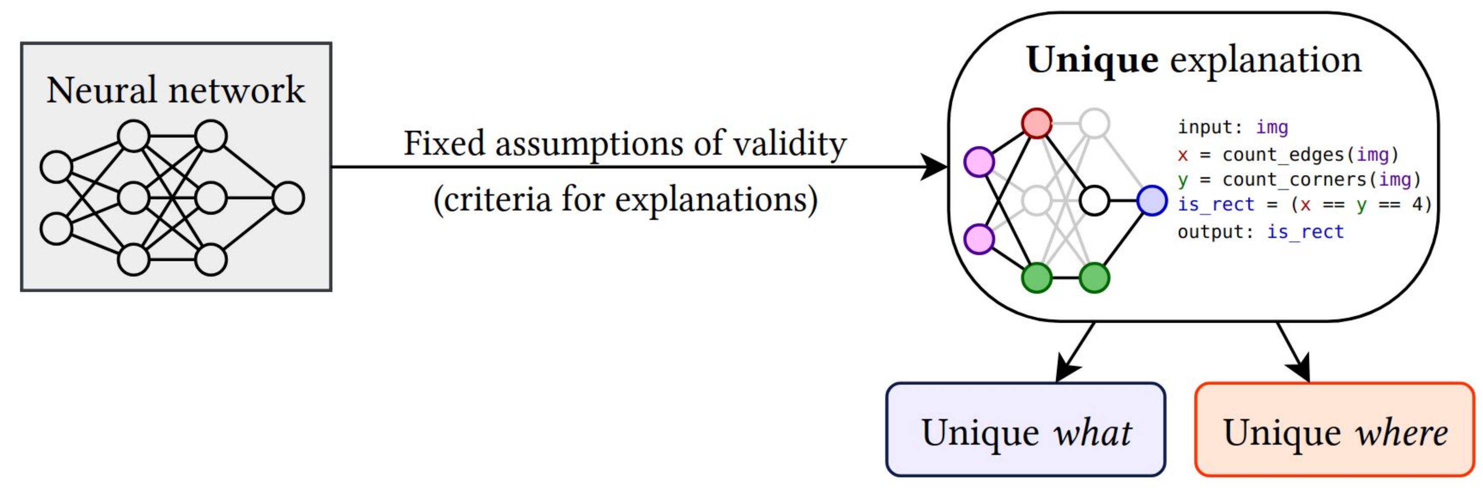
Intervention Interchange Accuracy (IIA)[1], measures the causal alignment of a *(mapping, algorithm)* pair, by performing counterfactual interventions on the variables of the high-level algorithm and those of the model through the mapping.



[1] Geiger, Atticus; Zhengxuan Wu; Hanson Lu; Josh Rozner; Elisa Kreiss; Thomas Icard; Noah D. Goodman; and Christopher Potts. 2022. *Inducing causal structure for interpretable neural networks.* In Proceedings of ICML.

Research question

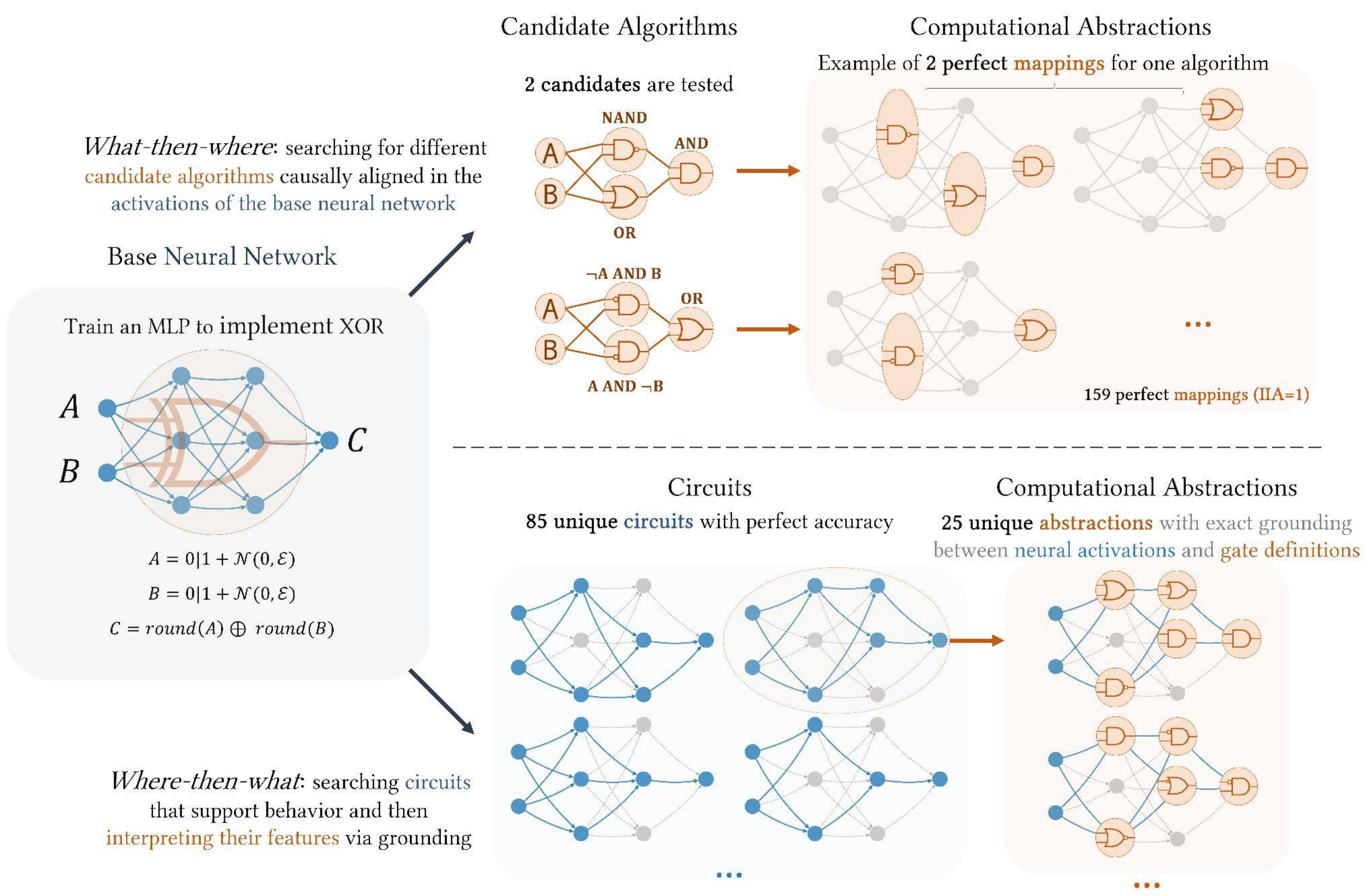
We ask whether MI is **identifiable**, borrowing this concept from statistics: When using fixed criteria for explaining a model's behavior, is the explanation unique? Is the *where* unique? Is the *what* unique?



Setup and results

- We train miniature multi-layer perceptrons (MLPs) on Boolean functions (XOR).
- We search for Boolean circuit explanations:
  - What sequence of logic gates is implemented by the MLP?
  - Where in the network is each gate implemented?

We **exhaustively** enumerate candidate algorithms and mappings, and test them with existing criteria (circuit error and mapping consistency for *where-then-what*, and IIA for *what-then-where*).



Even with strict causal alignment methods, we find multiple, incompatible explanations of the same neural computation. We encounter identifiability failures at every stage:

The *what* is not unique → Given the *what*, the *where* is not unique

The *where* is not unique → Given the *where*, the *what* is not unique

Additionally, the problem does not disappear when increasing the size of the network or changing training dynamics (duration, noise, multi-task setting).

Discussion and future work

Where to go from here? We suggest possible paths forward:

Change the criteria: Refine validity criteria with stronger constraints (based on causal abstraction), or use multi-criteria validation for explanations.

Change the expectation: Depending on the pragmatic goals of interpretability, uniqueness may not be required for predictability or controllability. However, if interpretability is expected to provide understanding, then non-identifiability becomes a problem.

Fundamental limits? In some scientific domains, multiple valid theories coexist; MI may similarly be underdetermined, and uniqueness might be unachievable without additional constraints.

ICLR 2025 — {maxime.meloux, francois.portet, silviu.maniu, maxime.peyrard}@univ-grenoble-alpes.fr