Towards Compositional Interpretability for XAI

Sean Tull Robin Lorenz Stephen Clark
Ilyas Khan Bob Coecke

QUANTINUUM

Applied Category Theory 2024, University of Oxford

Motivation

Today's AI models lack **interpretability**, which is a major safety concern in **high-stakes** areas (e.g. finance, health).

How does the model work?

Is it biased?

Why was the output X and not Y?

Motivation

Today's AI models lack **interpretability**, which is a major safety concern in **high-stakes** areas (e.g. finance, health).

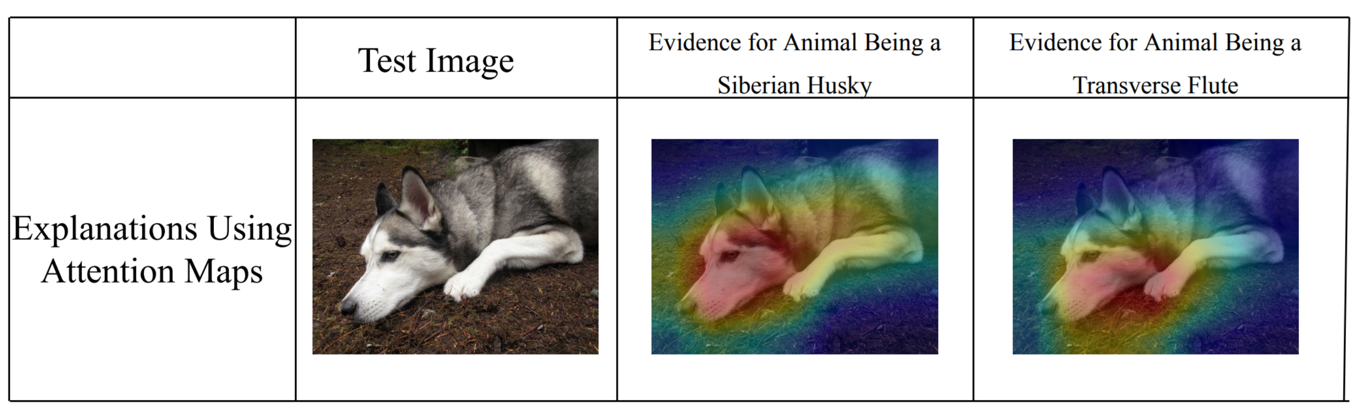
How does the model work?

Is it biased?

Why was the output X and not Y?

eXplainable Artificial Intelligence (XAI) hopes to solve this, focusing on **post-hoc explanations** for outputs (e.g. counterfactual explanations, salience maps).

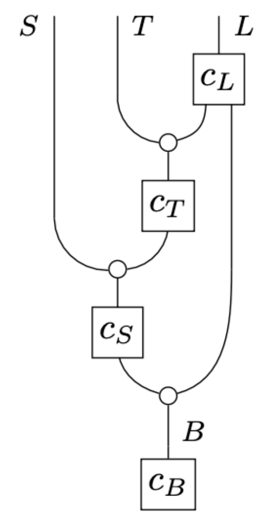
- However, these methods been **criticised** (e.g. Rudin 2019)
- There is no standard definition of 'interpretability' or 'explanation'.

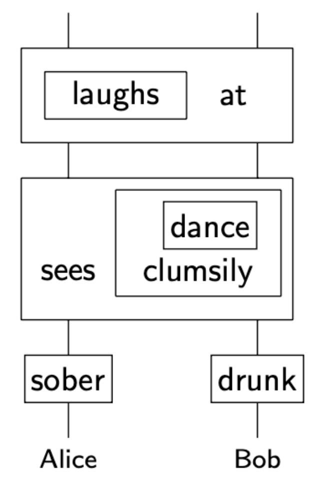


Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead,* 2019.

Motivation

Intuition: compositionally structured models are more interpretable.





How to make this precise? Aren't neural networks compositional?

This work:

- Gives a compositional formalism for defining AI models and interpretability
- Studies how compositional structure can give explainable models

Compositional Models

A monoidal **signature** G consists of sets:

- G_{ob} of 'objects' (**variables)**
- G_{mor} of 'morphisms' (**generators**), with lists of input and output variables
- optional equations.

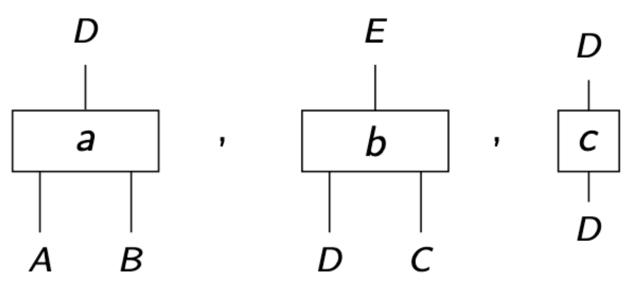
A **compositional model** $\mathbb{M}$ is given by a functor:

$$\mathbb{S} = \mathbf{Free}(G) \xrightarrow{\llbracket - \rrbracket} \mathbf{C}$$

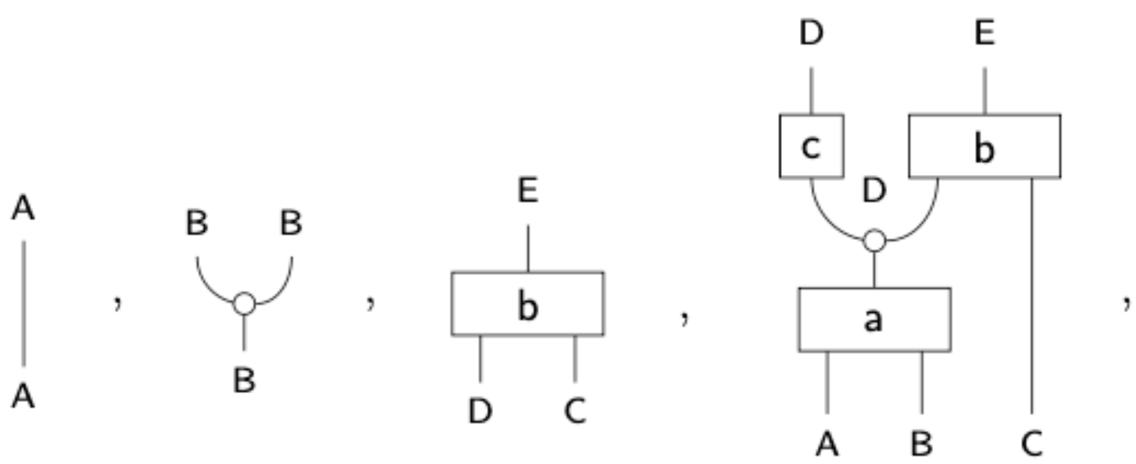
structure category of diagrams | **semantics** category

Example (language of cd-categories)

G



$\mathbb{S}$



Interpretations

An interpretation of a model consists of two aspects:

- an **abstract interpretation** $\mathcal{I}^{\mathsf{A}}$ interpreting variables and generators in G.

  e.g. V $\mapsto$ **'brightness'**

- a **concrete interpretation** $\mathcal{I}^{\mathsf{C}}$ interpreting morphisms in $\mathbf{C}$, such as states.

  e.g. 

$\mapsto$ **'dark'** 

$\mapsto$ **'bright'**

Interpretations

Formally, an **interpretation** of a model consists of:

- a signature $\mathbb{H}$ of '**human-friendly'** concepts
- partial maps of signatures $\mathcal{I}^{\mathsf{A}}$, $\mathcal{I}^{\mathsf{C}}$ such that the following commutes:

$$\begin{tikzcd} G \arrow[r, "{[\![-]\!]}"] \arrow[dr, "\mathcal{I}^{\mathsf{A}}"'] & \mathbf{C}_G \arrow[d, "\mathcal{I}^{\mathsf{C}}"] \\ & \mathbb{H} \end{tikzcd}$$

abstract

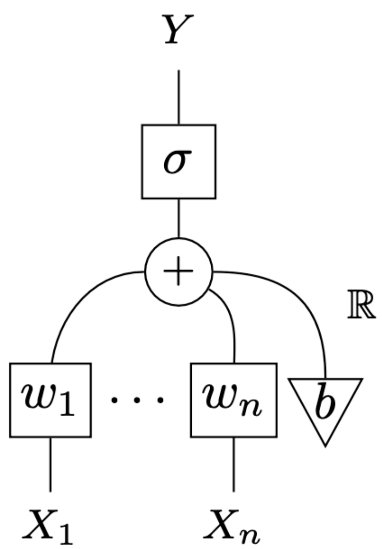
concrete

Variable V **has a concrete interpretation** when $I^C(v)$ is defined for all $v : I \to V$ in $\mathbf{C}$.

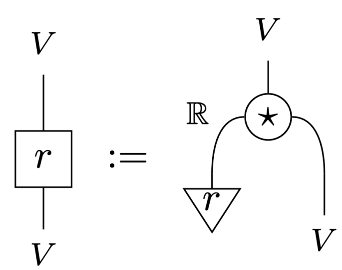
Here $\mathbf{C}_G$ has as objects lists of variables $(\mathsf{A}_i)_{i=1}^n$ and as morphisms $f : (\mathsf{A}_i)_{i=1}^n \to (\mathsf{B}_j)_{j=1}^m$ those $f : \otimes_{i=1}^n A_i \to \otimes_{j=1}^m B_j$ in $\mathbf{C}$.

Neural Networks

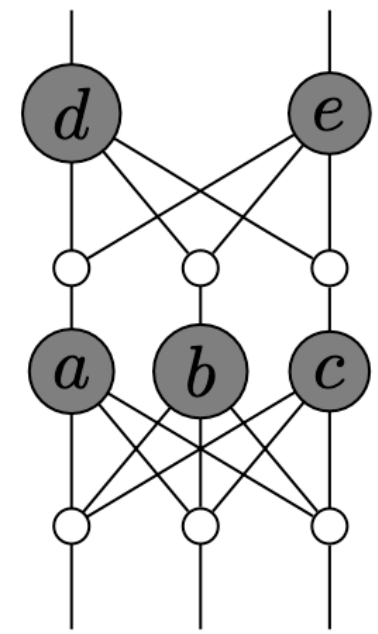
Neuron



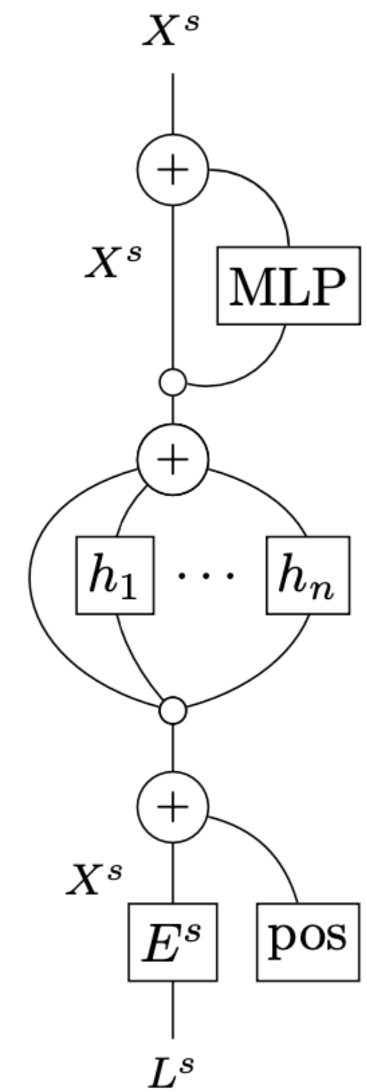
where



Network



Transformer



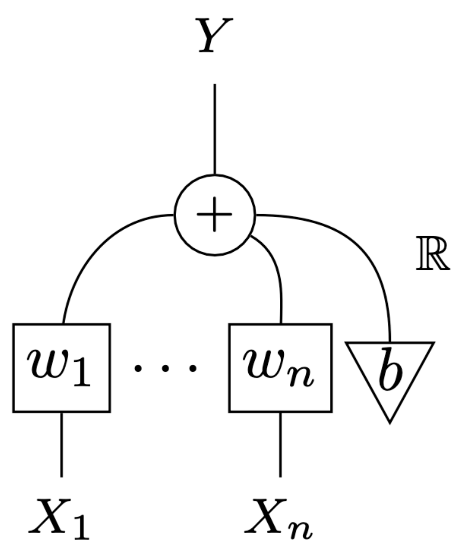
In the category $\mathbf{NN}$ of functions $f\colon \mathbb{R}^n \to \mathbb{R}^m$.

Observations

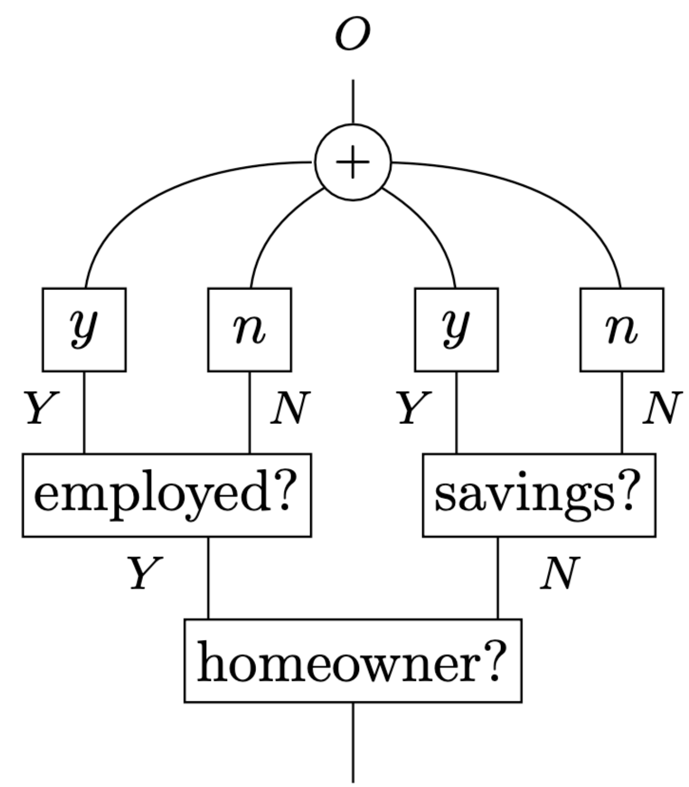
- Some forms of composition are common in ML.
- Compositional structure $\not\Rightarrow$ interpretability
- Only inputs and outputs typically interpretable, so this is where XAI focuses

Intrinsically Interpretable models

Linear



Rule-based



in $\mathbf{Set}$.

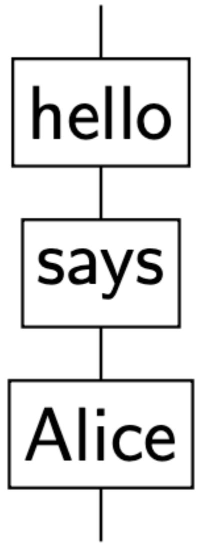
Observations

- **'Intrinsic interpretability'** is manifest diagrammatically. (The way in which the model is interpretable matches its diagram).

Compositionally Interpretable Models

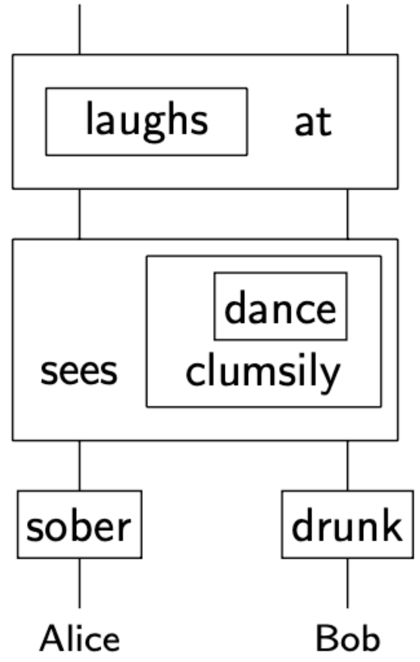
We call a model $\mathbb{M}$ **compositionally interpretable (CI)** when it has a complete abstract interpretation.

These include intrinsically interpretable models, and:

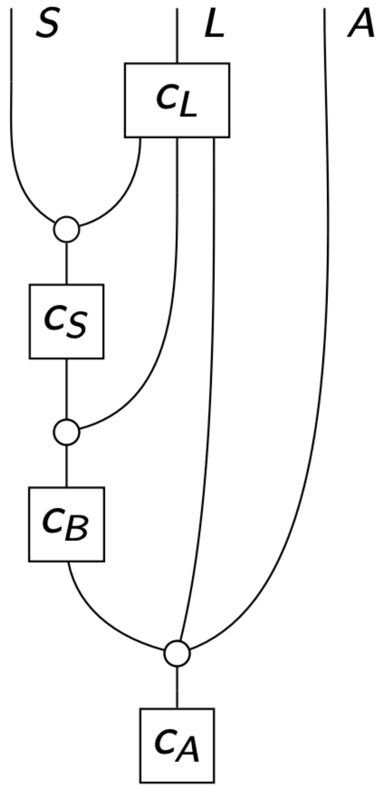


RNN

in $\mathbf{NN}$.



DisCoCirc



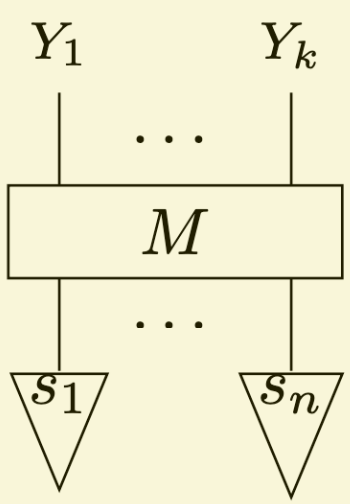
Causal models in $\mathbf{Mat}_{\mathbb{R}}^{+}$.

Studied in both ACT and **Causal ML.**

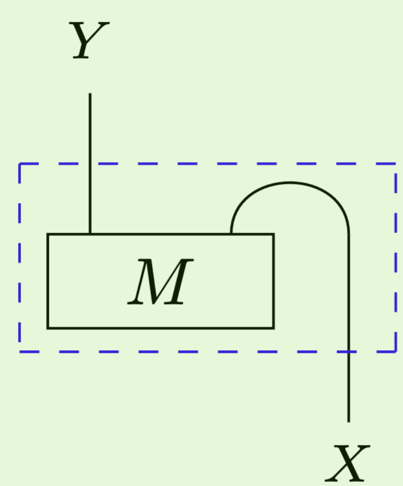
Compositional Frameworks

One way to capture how ‘rich’ the compositional structure is to consider its **framework**: what meaningful processes does it let us construct?

Input-output models

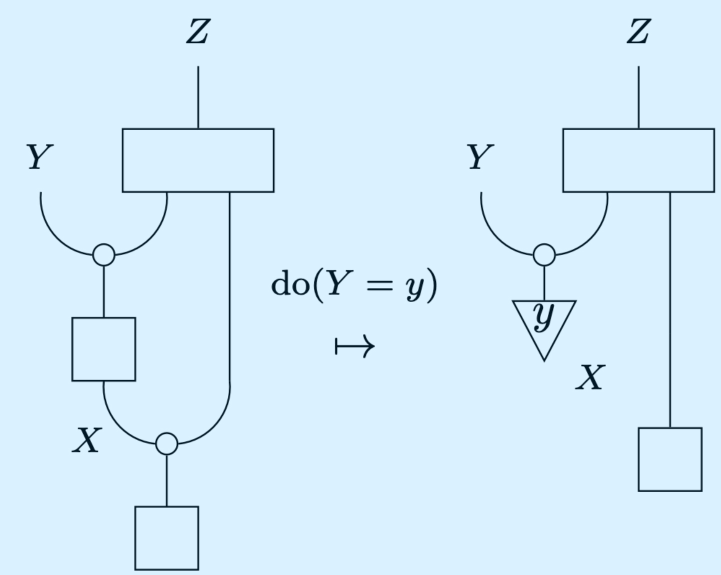


Statistical models

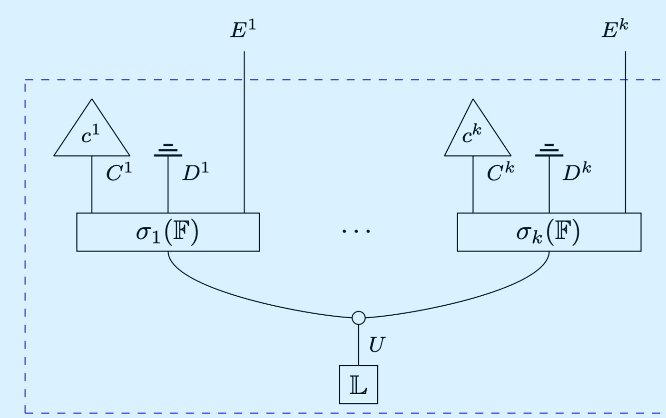


Conditionals P(Y|X)

Causal models



Interventions



Counterfactuals

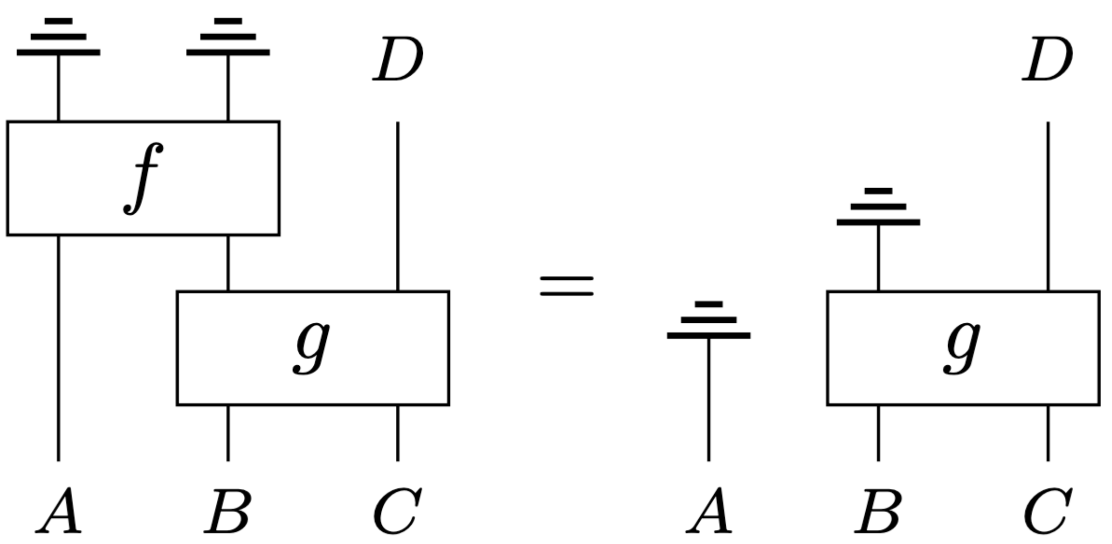
Explanations from Diagrams

How exactly does the compositional structure of a CI model yield **explanations** for its behaviour?

We propose three ways which are purely **diagrammatic**, and so in particular apply equally to e.g. classical or **quantum** models.

Influence Relations

For models based on (discard-preserving) **channels**, the explicit structure of a diagram lets us see which inputs can **influence** (or **signal to**) which outputs.

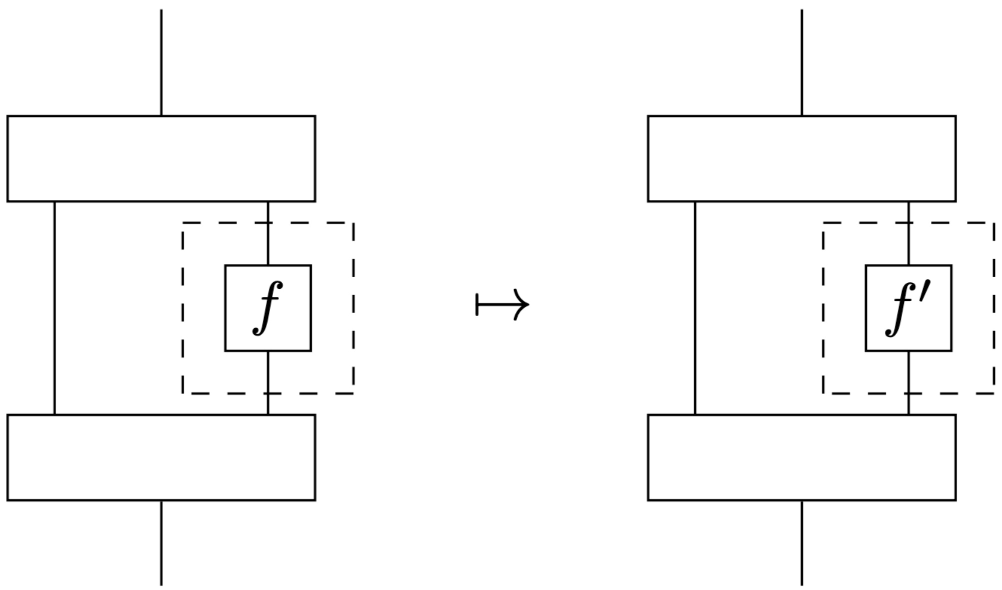


This is not possible for trivial compositional structure **e.g.** fully-connected NNs.

Diagram Surgery

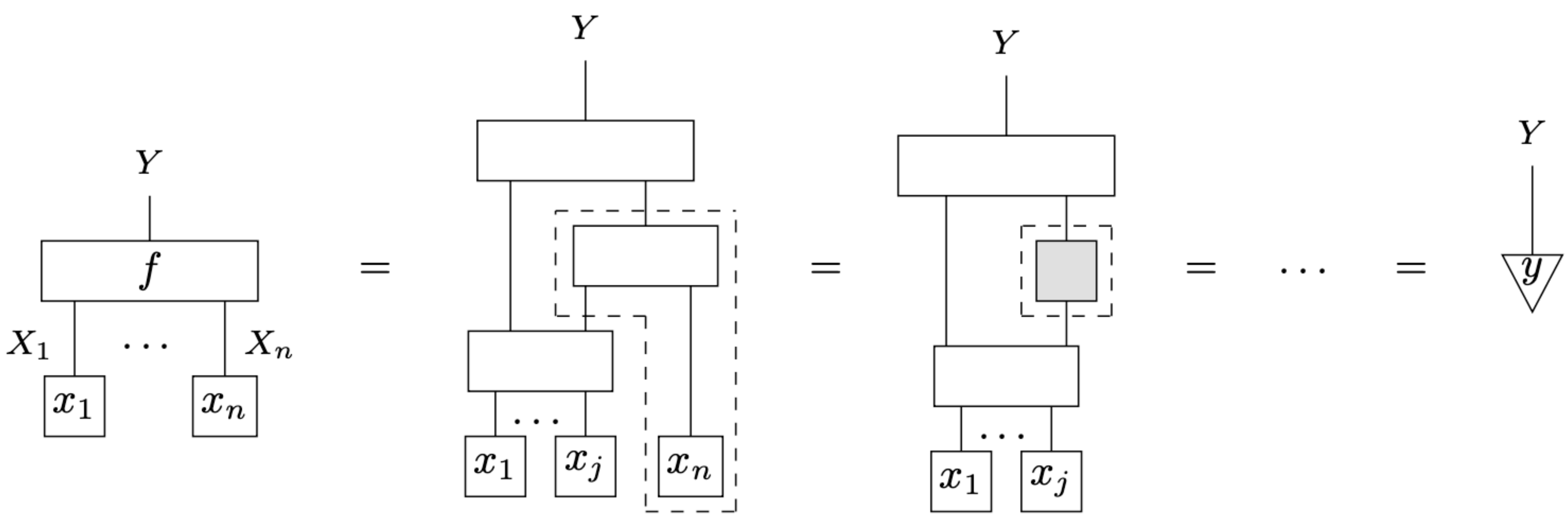
Each piece of an interpreted diagram forms a point we can *intervene* on by **diagram surgery**, to learn more about the process.

Generalises causal interventions, and CFEs to internal components.



Rewrite Explanations

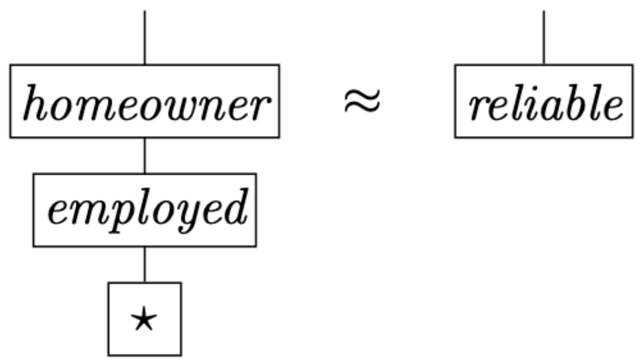
A **rewrite explanation** of an (approximate) equality $D = D'$ between interpreted diagrams consists of a collection of further such equations $(D_i = D_i')_{i=1}^{n}$ and a rewrite proof that these imply $D = D'$.

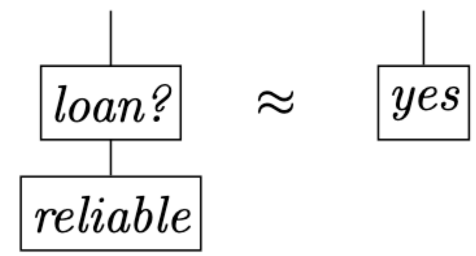


To count as an **explanation**, all diagrams involved must be interpreted.

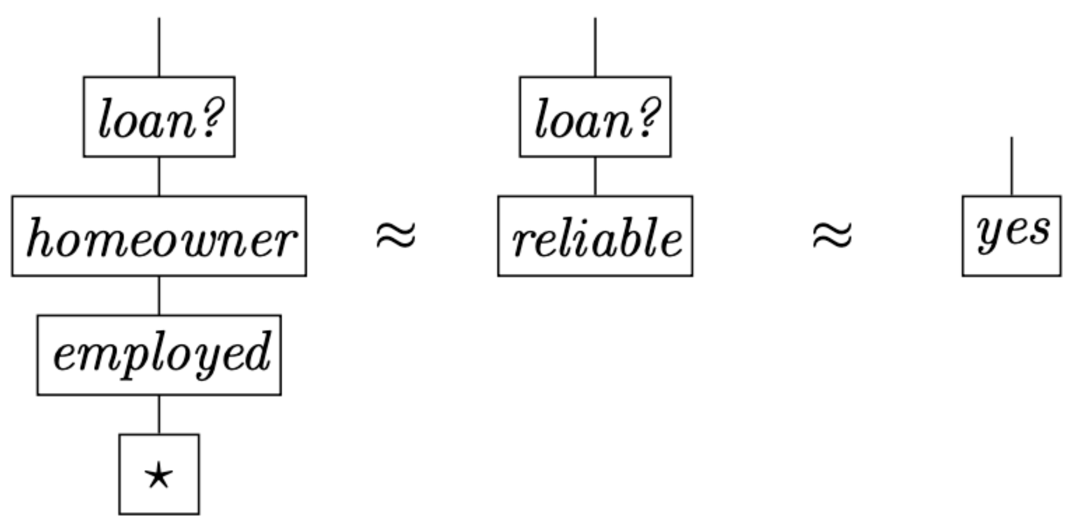
Rewrite Explanations

Suppose a bank uses an RNN model, which (almost) always grants an employed homeowner a loan. An explanation is given by approximate equalities:



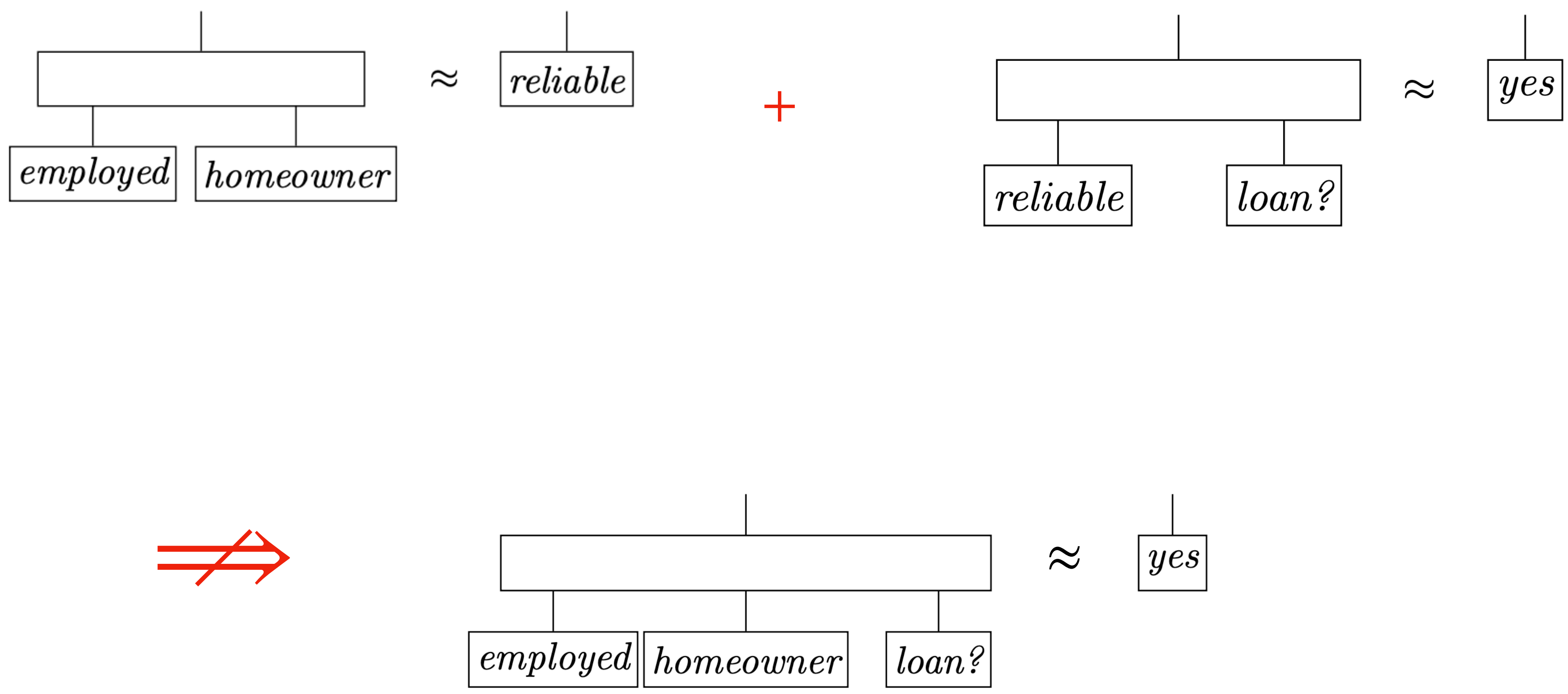


and the proof:



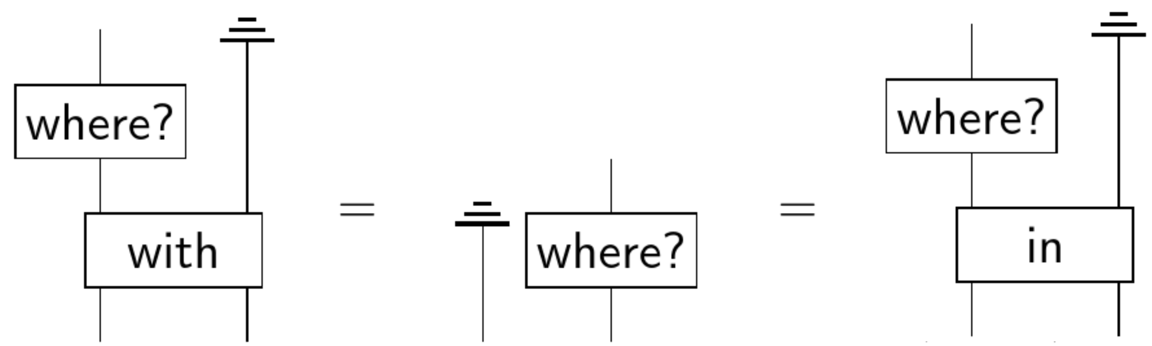
Rewrite Explanations

Such an argument is not possible for a black-box NLP model (e.g transformer):

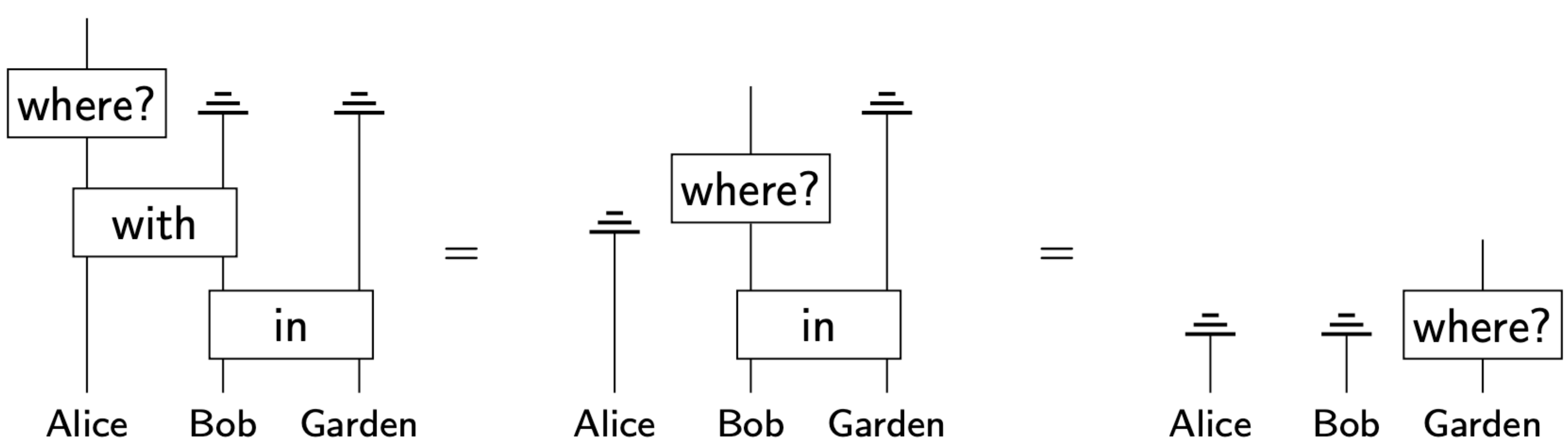


Rewrite Explanations

A DisCoCirc model of text '*Alice is with Bob. Bob is in the garden. Where is Alice?*'. Suppose the following equations hold:



A rewrite explanation for the answer *'garden'* could take the form:



Outlook

- Compositional approach natural for defining AI models and interpretability
- Leads to considering **compositionally interpretable (CI)** models
- These allow **diagrammatic explanations** for their behaviour

Outlook

- Compositional approach natural for defining AI models and interpretability
- Leads to considering **compositionally interpretable (CI)** models
- These allow **diagrammatic explanations** for their behaviour

Future directions:

Upgrading **rewrite explanations** as an XAI tool: where do the equations come from?

Finding more kinds of CI models

How do we **learn compositional structure**? (cf causal representation learning)

Explore further ways to relate NNs to a model, e.g. **causal abstraction**

Outlook

- Compositional approach natural for defining AI models and interpretability
- Leads to considering **compositionally interpretable (CI)** models
- These allow **diagrammatic explanations** for their behaviour

Future directions:

Upgrading **rewrite explanations** as an XAI tool: where do the equations come from?

Finding more kinds of CI models

How do we **learn compositional structure**? (cf causal representation learning)

Explore further ways to relate NNs to a model, e.g. **causal abstraction**

Thanks!

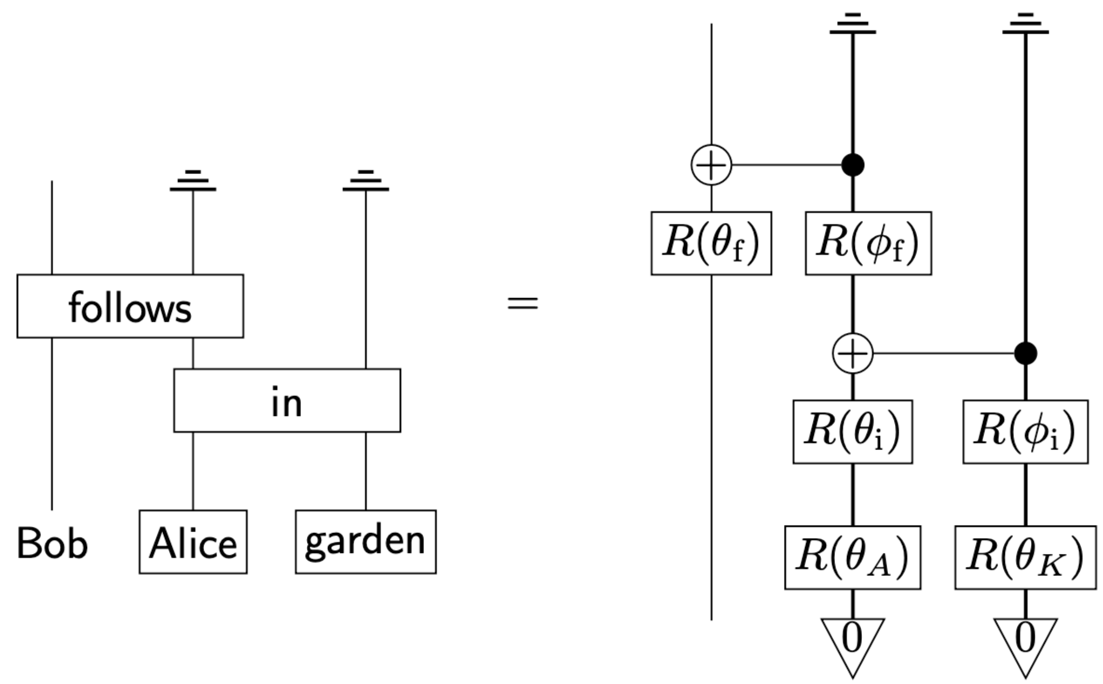
Bonus: Functorial Interpretations

$$
\begin{tikzcd}
\mathbb{S} \arrow[r, "\mathcal{I}^{\mathsf{A}}"] \arrow[dr, "\llbracket - \rrbracket"'] & \mathbb{H} \\
& \mathbf{C}_G \arrow[u, "\mathcal{I}^{\mathsf{C}}"']
\end{tikzcd}
$$

Quantum Models

A categorical treatment of interpretability is natural for **quantum AI models** since:

- Is model-agnostic so can compare classical vs quantum
- Quantum models are defined compositionally, as **circuits**



The notion of CI, and our explanation techniques, **apply equally** to quantum models.

Diagram Surgery

Each piece of an interpreted diagram forms a point we can *intervene* on by **diagram surgery**, to learn more about the process.

Generalises causal interventions, and CFEs to internal components.