Gluing Explanations in Projective Interpretability

Paul Wang

LIP6, Sorbonne Université

Multi-Agent Systems Team Seminar
April 2026

Outline

The interpretability problem

Interpretability = describe/explain a model's behaviour in human-understandable terms.

Most explanations are developed and validated on a **specific subset of the data**:

Probes

trained on
a specific
data distribution

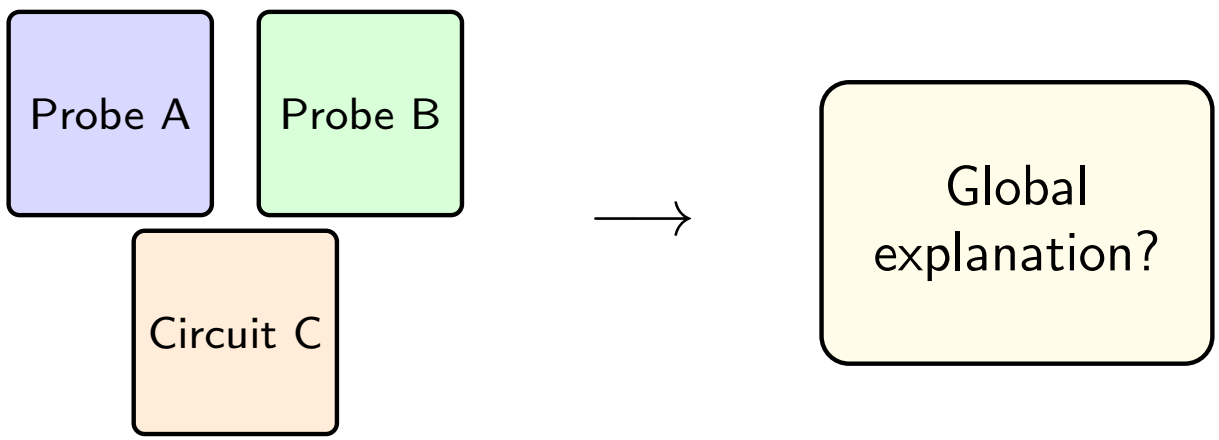
Circuits

identified
for specific
input patterns

The central question

Question

If I have explanations that each work on a **region of the data**, **do they fit together** into a coherent global picture?



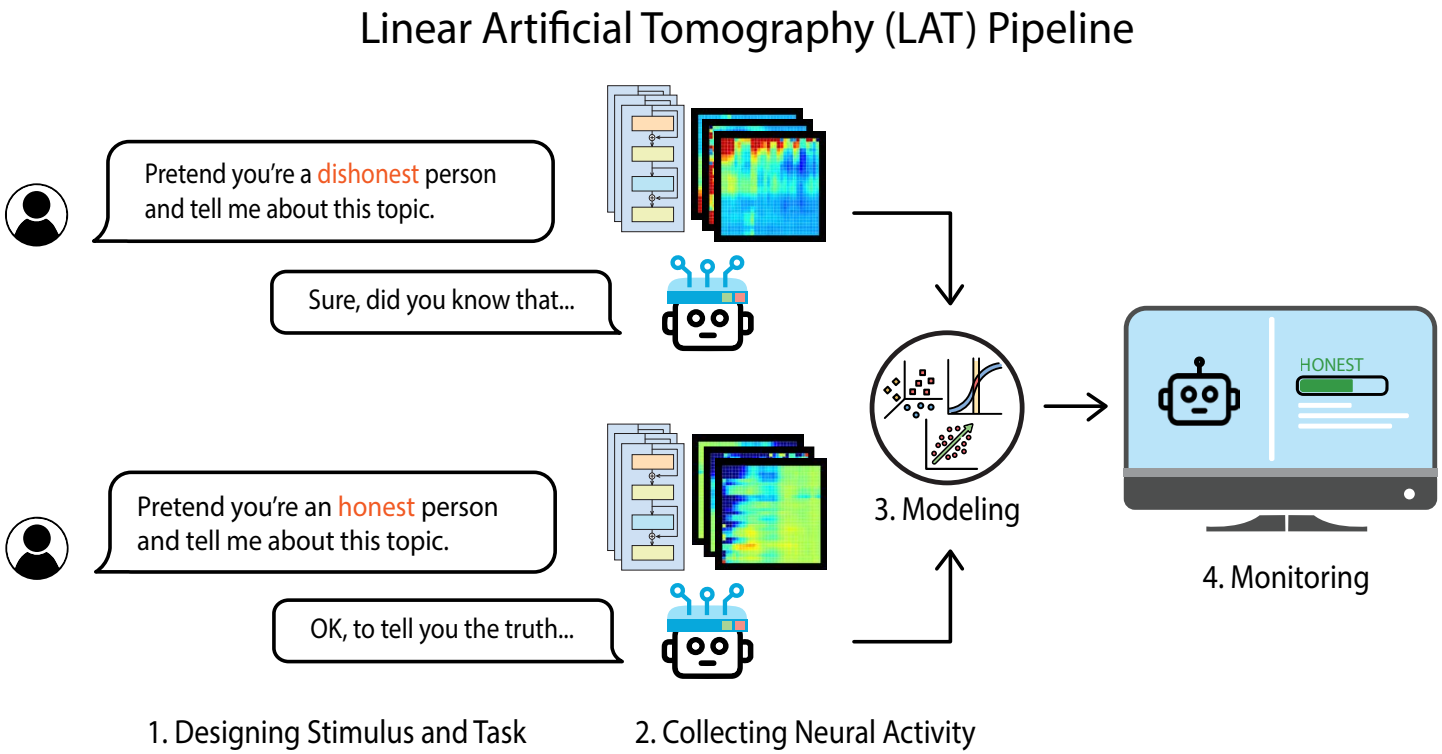
Let's make this precise!

Outline

Outline

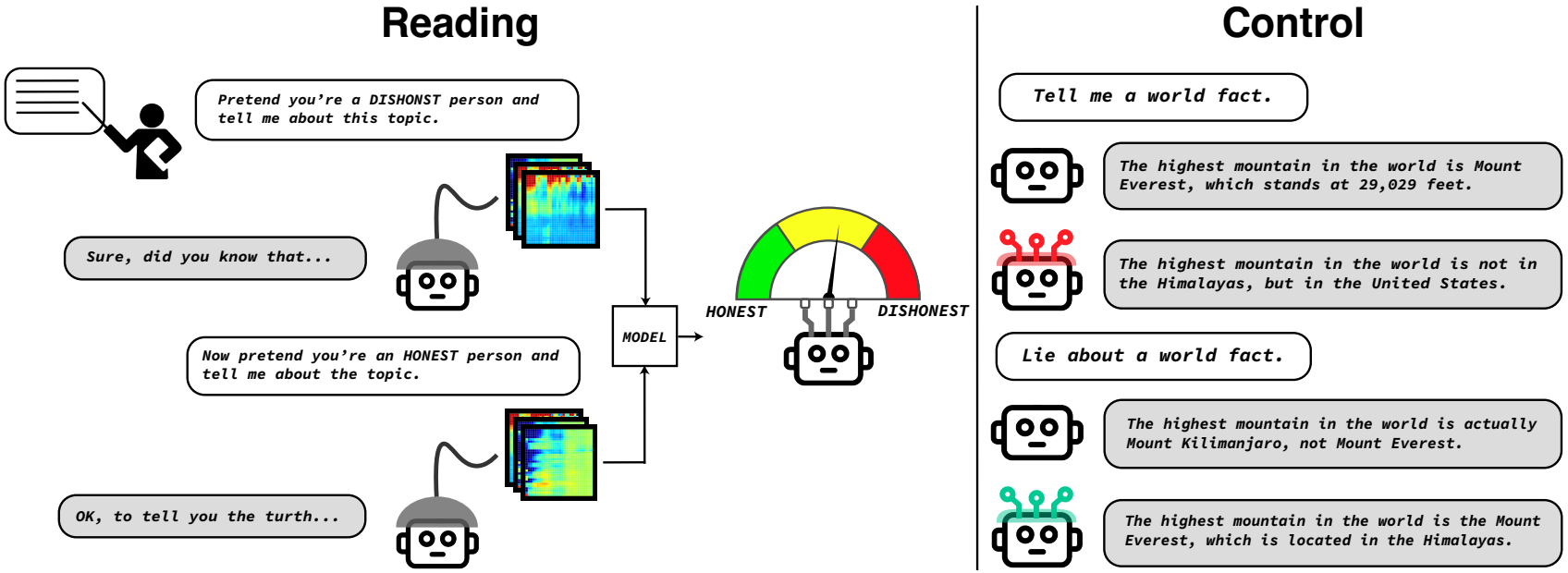
Representation Engineering

Key idea [Zou+23]: high-level concepts (sentiment, truthfulness, safety) are encoded as directions in activation space.



This is *a priori* data-specific.

Representation Engineering



A **probe** $v^\top : \mathbb{R}^d \to \mathbb{R}$ *reads off* a concept score from the hidden state. A family of probes $(v_1, \ldots, v_k)$ yields a map $\mathbb{R}^d \to \mathbb{R}^k$.

Probes as coordinate projections

A family of probes **projects** the model's high-dimensional state/input/output onto a – hopefully more interpretable – high-level summary.

This is a priori a local operation:

- Trained on a specific **subset of data**
- Valid for a specific **region of activation space**

Questions

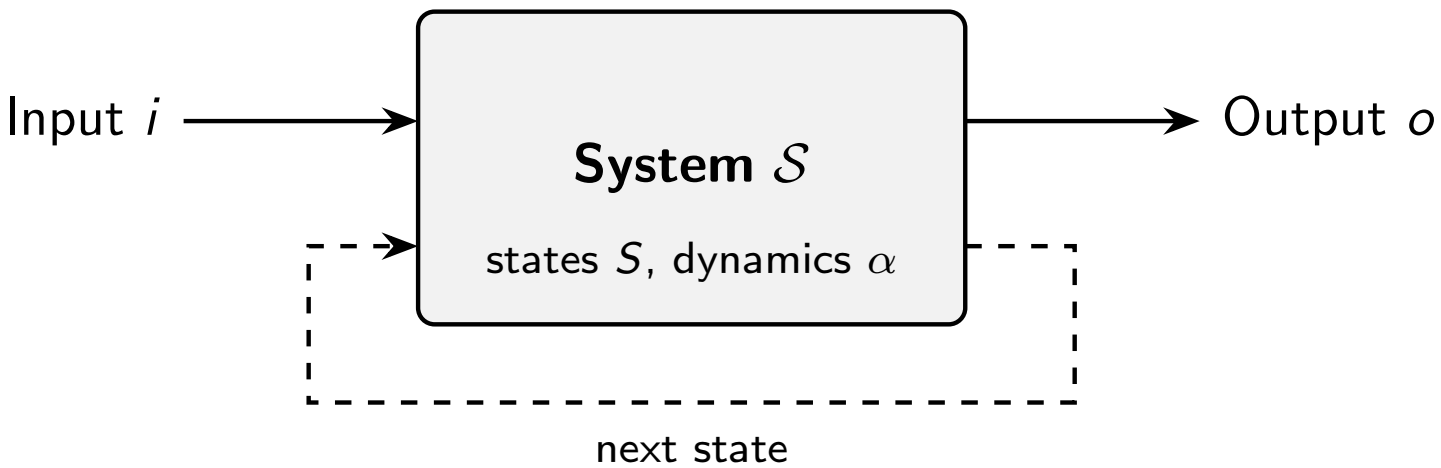
What kinds of explanations can we build using probes? Given probe-based explanations that each work on their **own region of the data**, do they combine into a single global explanation on the union?

Outline

Formalizing: systems

A **system** (in our context) has inputs, outputs, possibly internal states, and dynamics:

$$\alpha : S \times I \to S \times O$$



Examples: a neural network layer, a finite automaton, a Markov decision process. In many cases, S is trivial (singleton).

A **restricted system** is given by subsets of inputs/states.

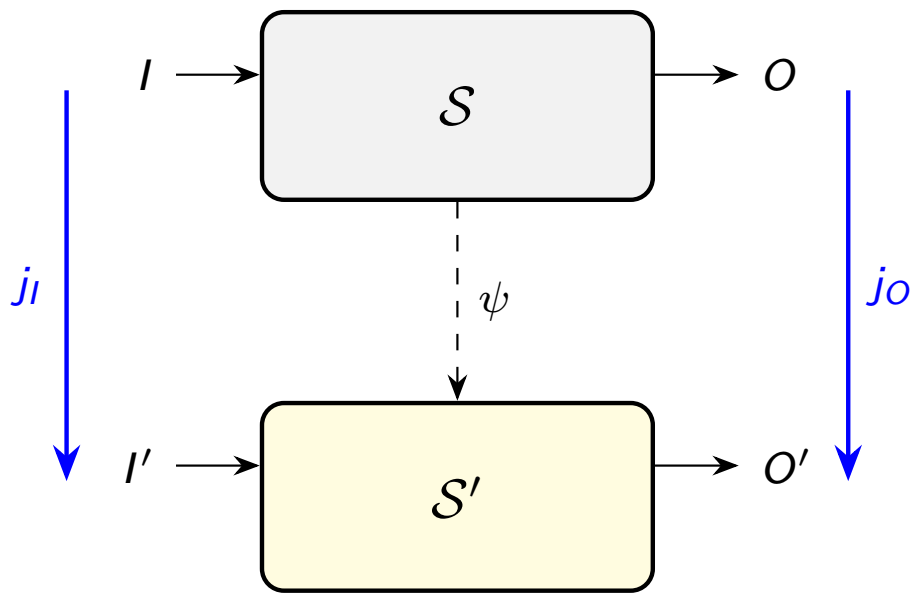
Formalizing: explanations

A **judge** $j = (j_I, j_O)$ specifies the interpretation at I/O level:

- $j_I : I \to I'$ groups inputs that "look the same" to the interpreter
- $j_O : O \to O'$ groups outputs similarly

One can construct such using probes from an auxiliary model.

An **explanation** is a system $\mathcal{S}'$ with interface (I', O'), along with a representation map $\psi : S \to S'$ that is compatible with the original dynamics through the judge:



The dynamics square

The key constraint: both paths through this diagram must agree.

$$\begin{array}{ccc} S \times I & \xrightarrow{\alpha} & S \times O \\ {\scriptstyle \psi \times j_I}\downarrow & & \downarrow{\scriptstyle \psi \times j_O} \\ S' \times I' & \xrightarrow{\alpha'} & S' \times O' \end{array}$$

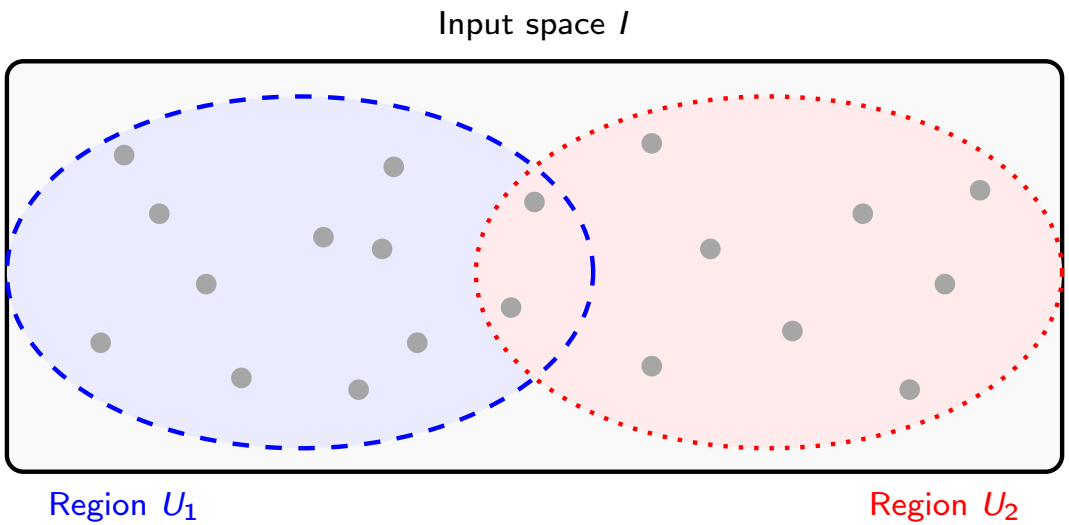
Translate then run dynamics = run dynamics then translate.

The explanation doesn't just describe the current state or input: it must give a high-level description of the dynamics – this is specified in the system $\mathcal{S}'$.

Outline

Restricting to data regions

In practice, we don't explain the *whole* model at once. We look at **subsets of the data**:



A **covering**: regions $\{U_1, U_2, \ldots\}$ that together cover all the input space (overlap allowed).

An explanation on U_i is a system that reproduces the dynamics **on that region**, i.e. an explanation of $\mathcal{S}|_{U_i}$.

Locality of explanations

For each region U, we collect all possible explanations of the *restricted system* $\mathcal{S}|_U$ into a set $\tilde{\mathcal{F}}_j(U)$.

When one region is contained in another ($U' \subseteq U$), we can **restrict**: an explanation on U gives one on U' by forgetting the extra data.

Locality of explanations

For each region U, we collect all possible explanations of the *restricted system* $\mathcal{S}|_U$ into a set $\tilde{\mathcal{F}}_j(U)$.

When one region is contained in another ($U' \subseteq U$), we can **restrict**: an explanation on U gives one on U' by forgetting the extra data.

This "regions $\to$ sets of explanations" assignment, with restriction maps, is a **presheaf**.

Two key questions

- **Separation:** if two explanations agree on every piece, do they agree globally?
- **Gluing:** if local explanations are compatible on overlaps, can they be assembled into a global one?

A presheaf where both hold is called a **sheaf**.

Beyond exact explanations

In practice, exact compatibility is too much to ask:

- Probes have nonzero training loss
- Models may involve stochastic components (sampling, dropout)
- Explanations simplify by design — they *approximate*

Beyond exact explanations

In practice, exact compatibility is too much to ask:

- Probes have nonzero training loss
- Models may involve stochastic components (sampling, dropout)
- Explanations simplify by design — they *approximate*

We relax the dynamics square to commute **up to** ε:

$$d\big((\psi \times j_O) \circ \alpha,\ \alpha' \circ (\psi \times j_I)\big) \leq \varepsilon$$

This gives ε-**explanations**: a richer presheaf $\tilde{\mathcal{F}}_j^{\varepsilon} \supseteq \tilde{\mathcal{F}}_j$.

Does pairwise compatibility of ε-explanations imply global compatibility?

Outline

Helly's theorem: output dimension matters

Theorem

For ε-explanations: checking $(k+1)$-wise compatibility **suffices** for global compatibility, *for stateless* $\mathcal{S}'$, *when* $O' \subseteq \mathbb{R}^k$ *is closed convex*.

Caveat: this does not guarantee *continuity* of the explanatory system.

Helly's theorem: output dimension matters

Theorem

For ε-explanations: checking $(k+1)$-wise compatibility **suffices** for global compatibility, *for stateless* $\mathcal{S}'$, *when* $O' \subseteq \mathbb{R}^k$ *is closed convex*.

Caveat: this does not guarantee *continuity* of the explanatory system.

For the proof, we use:

Theorem (Helly [Hel23])

In $\mathbb{R}^k$: if any $k+1$ sets in a given family of convex sets intersect, then the **global** intersection of that family is nonempty.

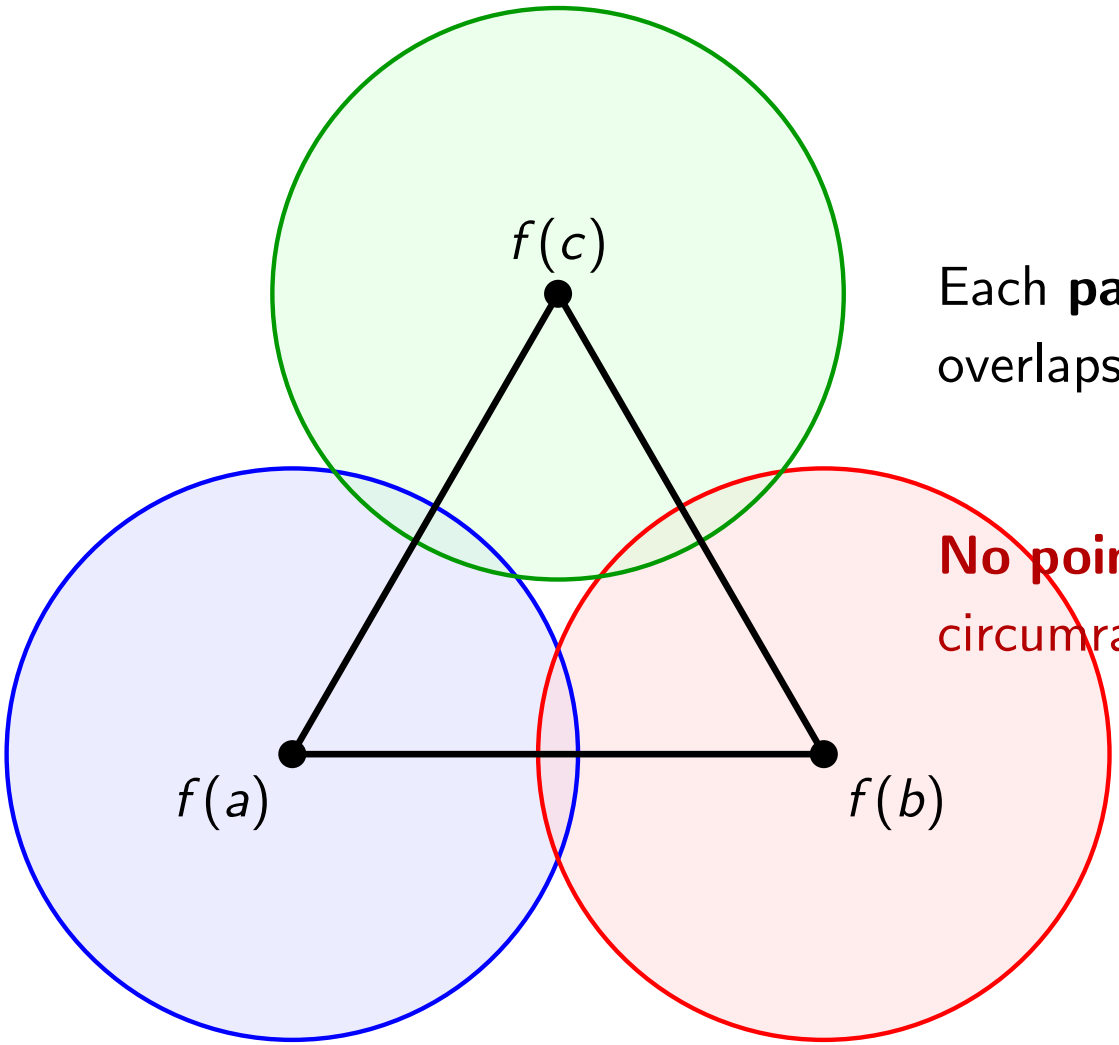
Helly's theorem: output dimension matters

Output type	Must check	Pairwise enough?
Discrete *exact* classification	2-wise *exact*	✓
k-class soft (simplex Δ^{k-1})	k-wise	only if $k = 2$
Regression ($\mathbb{R}^k$)	$(k+1)$-wise	only if $k = 1$

Takeaway: output dimension $\dim(O')$ is a complexity parameter for explanation gluing. Richer outputs need more regions checked simultaneously.

Abstract example: the triangle obstruction

Output in $\mathbb{R}^2$. Three data regions each produce one output value, forming an equilateral triangle:



Each **pair** of ε-balls overlaps: midpoint within ε ✓

No point lies in all three: circumradius $> \varepsilon$ ×

Summary and open questions

1. Defined a notion of **local explanations**.
2. **Approximate explanations**: the framework extends to ε-tolerance. Pairwise reconcilability does not imply global reconcilability (triangle obstruction).
3. **Helly depth bound**: output dimension controls how many regions must be checked simultaneously. Binary: 2. Regression in $\mathbb{R}^k$: $k+1$.

Summary and open questions

1. Defined a notion of **local explanations**.
2. **Approximate explanations**: the framework extends to ε-tolerance. Pairwise reconcilability does not imply global reconcilability (triangle obstruction).
3. **Helly depth bound**: output dimension controls how many regions must be checked simultaneously. Binary: 2. Regression in $\mathbb{R}^k$: $k+1$.

Open questions:

- Can we **measure** the obstruction? (Cohomological invariants [EP08])
- How often does this happen in **real models**? (Experimental validation on transformers)

Thank you!

Questions?

References I

[AB11] Samson Abramsky and Adam Brandenburger. "The Sheaf-Theoretic Structure of Non-Locality and Contextuality". *New Journal of Physics* 13.11 (2011), p. 113036.

[EP08] Mário J. Edmundo and Nicholas J. Peatfield. "O-minimal Čech cohomology". *Quarterly Journal of Mathematics* 59.2 (2008), pp. 213–220.

[Hel23] Eduard Helly. "Über Mengen konvexer Körper mit gemeinschaftlichen Punkten". *Jahresbericht der Deutschen Mathematiker-Vereinigung* 32 (1923), pp. 175–176.

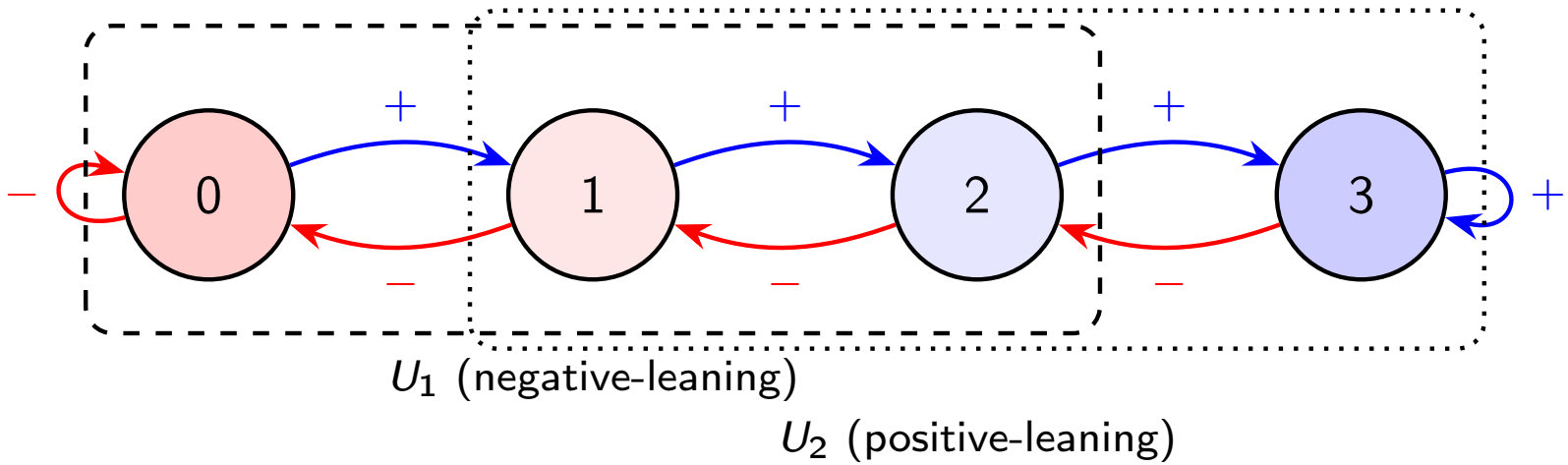
[MM94] Saunders Mac Lane and Ieke Moerdijk. *Sheaves in Geometry and Logic: A First Introduction to Topos Theory*. Universitext. Springer-Verlag, 1994.

References II

[Zou+23] Andy Zou et al. "Representation Engineering: A Top-Down Approach to AI Transparency". *arXiv preprint arXiv:2310.01405* (2023).

Bonus: the sentiment accumulator

A toy model of a sentiment classifier:



- States $\{0, 1, 2, 3\}$ = sentiment score; $I = \{+, -\}$. Judge: $j_O(0) = j_O(1) =$ "neg", $j_O(2) = j_O(3) =$ "pos"
- Each patch has a simple 2-state explanation. Compatible on overlap $\{1, 2\}$

Bonus: gluing failure in the sentiment accumulator

The problem: dynamics at the boundary.

At state 2 in the overlap:

- Input $+$ sends it to state 3 ("positive", output 1)
- Input $-$ sends it to state 1 ("negative", output 0)

The judge sees both inputs identically ($j_I = \mathrm{id}$), but the after-states reach **all four states** — requiring at least 4 explanatory states just to handle the overlap.

The obstruction

The dynamics at the boundary forces any global explanation to be as complex as the original system. The local simplification **cannot survive** globally.

This is a **gluing failure**: compatible local explanations that cannot be assembled.

Bonus: several notions of "same explanation"

Two explanations can agree in different senses:

- **Structural agreement**: the two share a common internal core (same states, same transitions on the shared part)
- **Behavioral agreement**: produce the **same outputs** for all input sequences — regardless of internal wiring
- **Restricted-interface agreement**: produce the same outputs **only for inputs reachable through the judge**

Structural $\Rightarrow$ behavioural $\Rightarrow$ restricted-interface.

Each notion leads to a different presheaf, and separation/gluing can differ.