Presheaves of Explanations on Categories of Systems

Paul Wang[1]

LIP6, Sorbonne Université

CMCS 2026

[1]Supported by a CoefficientGiving Grant.

The gluing question

Current interpretability methods produce data-**local** explanations:

- *Probes*: project hidden states onto interpretable concepts [Zou+23]
- *Automata extraction*: fit a finite-state machine to a subregion of behaviour [WGY18]

When multiple data-local explanations are available:

- **Consistency**: are the local explanations compatible on the overlaps?
- **Gluing**: do compatible local explanations assemble into a global one?

Analogy: Abramsky–Brandenburger [AB11]

Local measurement data fail to extend to a global section $\Rightarrow$ **contextuality**. Our situation is structurally analogous, but presheaves live on *categories of dynamical systems*.

Outline

Heterogeneous Mealy machines

A **heterogeneous Mealy machine**: $\mathcal{S} = (S_b, S_a, I, O, \alpha)$

$$\alpha : S_b \times I \to S_a \times O$$

- S_b = before-states, S_a = after-states (may differ)
- Heterogeneous systems arise from **restricting** to state subsets
- In $\mathbf{Set}^2$: coalgebras for $H(X, Y) = ((Y \times O)^I, \star)$
- Homogeneous case: $S_b = S_a =: S$. In **Set**: coalgebras for $H(X) = (X \times O)^I$

A **morphism** $\mathcal{S} \to \mathcal{S}'$: a 4-tuple (f_b, f_a, f_I, f_O) making the **dynamics square** commute:

$$\begin{array}{ccc} S_b \times I & \xrightarrow{\alpha} & S_a \times O \\ {\scriptstyle f_b \times f_I}\downarrow & & \downarrow{\scriptstyle f_a \times f_O} \\ S'_b \times I' & \xrightarrow[\alpha']{} & S'_a \times O' \end{array}$$

The site

Let $\mathcal{M}$ be the class of all monomorphisms in **Set**. An $\mathcal{M}$-**immersion** $\iota : \mathcal{S}_1 \hookrightarrow \mathcal{S}_2$: each component $\iota_b, \iota_a, \iota_I, \iota_O \in \mathcal{M}$.

Fix a system $\mathcal{S}$. The objects of the site are $\mathcal{M}$-immersions $\mathcal{S}_U \hookrightarrow \mathcal{S}$. The morphisms are morphisms of $\mathcal{M}$-immersions.

Coverings: finite families $\{\iota_\alpha : \mathcal{S}_\alpha \hookrightarrow \mathcal{S}_U\}$ jointly surjective on $S_{U,b} \times I_U$ and $S_{U,a} \times O_U$.

This defines a Grothendieck pretopology $(\mathrm{Sys}_{\mathcal{M}}, \mathrm{Cov})$. Pullbacks $=$ componentwise intersections.

Judges and the presheaf hierarchy

A **judge** $j = (j_I, j_O)$: maps $j_I : I \to I'$, $j_O : O \to O'$ to an *interpretable interface*. In practice: j_I and j_O are *feature maps*.
A **section** over $\mathcal{S}_U \hookrightarrow \mathcal{S}$ of the presheaf $\tilde{\mathcal{F}}_j$: a system $\mathcal{S}' = (S', I', O', \alpha')$ with a morphism $\psi : \mathcal{S}_U \to \mathcal{S}'$ through the judge j.

Judges and the presheaf hierarchy

A **judge** $j = (j_I, j_O)$: maps $j_I : I \to I'$, $j_O : O \to O'$ to an *interpretable interface*. In practice: j_I and j_O are *feature maps*.
A **section** over $\mathcal{S}_U \hookrightarrow \mathcal{S}$ of the presheaf $\tilde{\mathcal{F}}_j$: a system $\mathcal{S}' = (S', I', O', \alpha')$ with a morphism $\psi : \mathcal{S}_U \to \mathcal{S}'$ through the judge j.

Four quotients of the raw presheaf, by decreasing fineness:

$$\tilde{\mathcal{F}}_j \twoheadrightarrow \mathcal{F}_j^{\mathrm{cogerm}} \twoheadrightarrow \mathcal{F}_j^{\mathrm{beh}} \twoheadrightarrow \mathcal{F}_j^{\mathrm{ri}}$$

- $\tilde{\mathcal{F}}_j$: raw sections (too fine — remembers internal structure)
- $\mathcal{F}_j^{\mathrm{cogerm}}$: cogerm equivalence (common subsystem witness)
- $\mathcal{F}_j^{\mathrm{beh}}$: behavioural equivalence (same input-output trees, cf. Rutten [Rut00])
- $\mathcal{F}_j^{\mathrm{ri}}$: restricted-interface (sections with input space $j_I(I_U) \subseteq I'$, up to behavioural equivalence)

The landscape

Presheaf	**Separated?**	**Gluing?**
$\tilde{\mathcal{F}}_j$ (raw)	Yes	Yes
$\mathcal{F}_j^{\mathrm{cogerm}}$ (cogerm)	No	Yes
$\mathcal{F}_j^{\mathrm{beh}}$ (behavioural)	Yes	No
$\mathcal{F}_j^{\mathrm{ri}}$ (restricted-interface)	No	No

Caveat: Finicky, tweaking the definitions gives different results! Final versions might differ.

Adhesivity of systems categories

Theorem

Let $\mathcal{C}$ be adhesive with finite products. Then, $\mathrm{Sys}_{\mathrm{ho}}(I, O)$ is adhesive for all I, O in $\mathcal{C}$.

Key idea: The forgetful functor $U : \mathrm{Sys}_{\mathrm{ho}}(I, O) \to \mathcal{C}$ creates pullbacks and pushouts along monomorphisms.
Note: this does not require exponentials!

O-minimal structures

Working over **Set**: any injection is a morphism, many coverings.
No geometric control.

O-minimal structures

Working over **Set**: any injection is a morphism, many coverings.
No geometric control.

Definition [Dri98]

An **o-minimal structure** M expanding $(\mathbb{R}, <, +, \cdot)$: a sequence of Boolean algebras $\mathcal{D}_n \subseteq \mathcal{P}(\mathbb{R}^n)$, closed under products and projections, containing all semialgebraic sets, such that $\mathcal{D}_1$ consists **exactly** of finite unions of intervals and points.

Elements of $\mathcal{D}_n$ are **definable sets**; functions with definable graphs are **definable**.

Examples: $\mathbb{R}_{\mathrm{alg}}$ (semialgebraic), $\mathbb{R}_{\mathrm{an}}$ (subanalytic), $\mathbb{R}_{\mathrm{an,exp}}$ (adding exp).
Neural network activations (ReLU, sigmoid, softmax) are definable in $\mathbb{R}_{\mathrm{an,exp}}$ [JT20].

Cell decomposition

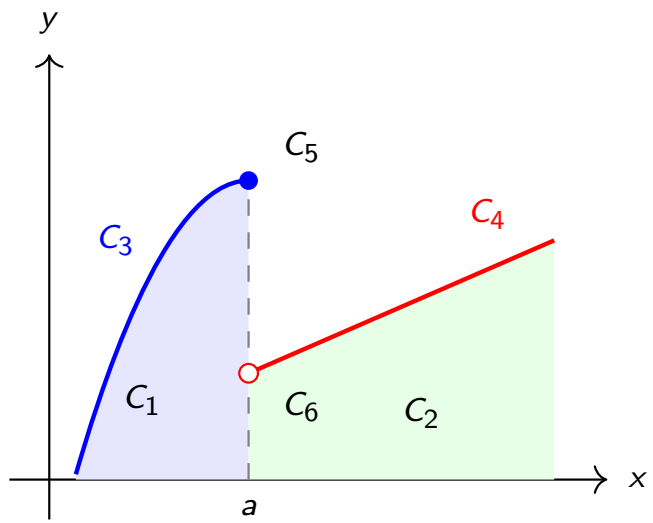
The axiom on $\mathcal{D}_1$ propagates to all dimensions:

Cell decomposition theorem

Every definable set $X \subseteq \mathbb{R}^n$ admits a **finite** partition into *cells*: connected definable pieces, each of controlled topological type.

Consequences:

- Finitely many connected components
- Well-defined dimension
- Piecewise monotonicity ($n = 1$)
- No space-filling curves



Closed subgraph of f (discontinuous at a): two 2-cells (C_1, C_2), two 1-cells (curves C_3, C_4), two 0-cells (C_5 closed, C_6 open).

O-minimality and dynamical systems

Why this matters for us:

1. $\mathrm{Def}(\mathrm{M}) \subset$ **Set** is still cartesian and **adhesive**
2. Definable open subsets give a **canonical class of morphisms** for our site
3. Finite cell decompositions $\Rightarrow$ **finite admissible covers** $\Rightarrow$ finite cohomology [EP08]
4. Tameness prevents pathological coverings (every covering has finitely many connected components)

$\mathcal{C} = \mathrm{Def}(\mathrm{M})$, $\mathcal{M} = \{\text{definable open immersions}\}$ is a useful instantiation: tame topology, existing sheaf-theoretic tools (e.g. constructible sheaves).

Summary and outlook

Summary.

- **Grothendieck sites** on Mealy machines
- **Several presheaves of explanations**: raw, cogerm, behavioural, and restricted-interface behavioural
- **Adhesivity theorem**: $\mathcal{C}$ adhesive $\Rightarrow \mathrm{Sys}_{\mathrm{ho}}(I, O)$ adhesive

Research Directions.

1. Cohomological obstructions: lack of global sections $\rightarrow$ non-trivial H^1?
2. Approximate (metric-enriched) versions, and/or continuity requirements ($Def_c(M)$).
3. Extensions for Markov categories?

Thank you!

Thank you!

paul.wang@lip6.fr

Paper: *Separation and Gluing of Explanations on Sites of Dynamical Systems*, [Wan26]

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.

[Dri98] Lou van den Dries. *Tame Topology and O-minimal Structures*. London Mathematical Society Lecture Note Series 248. Cambridge University Press, 1998.

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

[JNW96] André Joyal, Mogens Nielsen, and Glynn Winskel. "Bisimulation from Open Maps". *Information and Computation* 127.2 (1996), pp. 164–185.

[JT20] Ziwei Ji and Matus Telgarsky. "Directional convergence and alignment in deep learning". *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 17176–17186.

[LS05] Stephen Lack and Paweł Sobociński. "Adhesive and quasiadhesive categories". *RAIRO - Theoretical Informatics and Applications* 39.3 (2005), pp. 511–545.

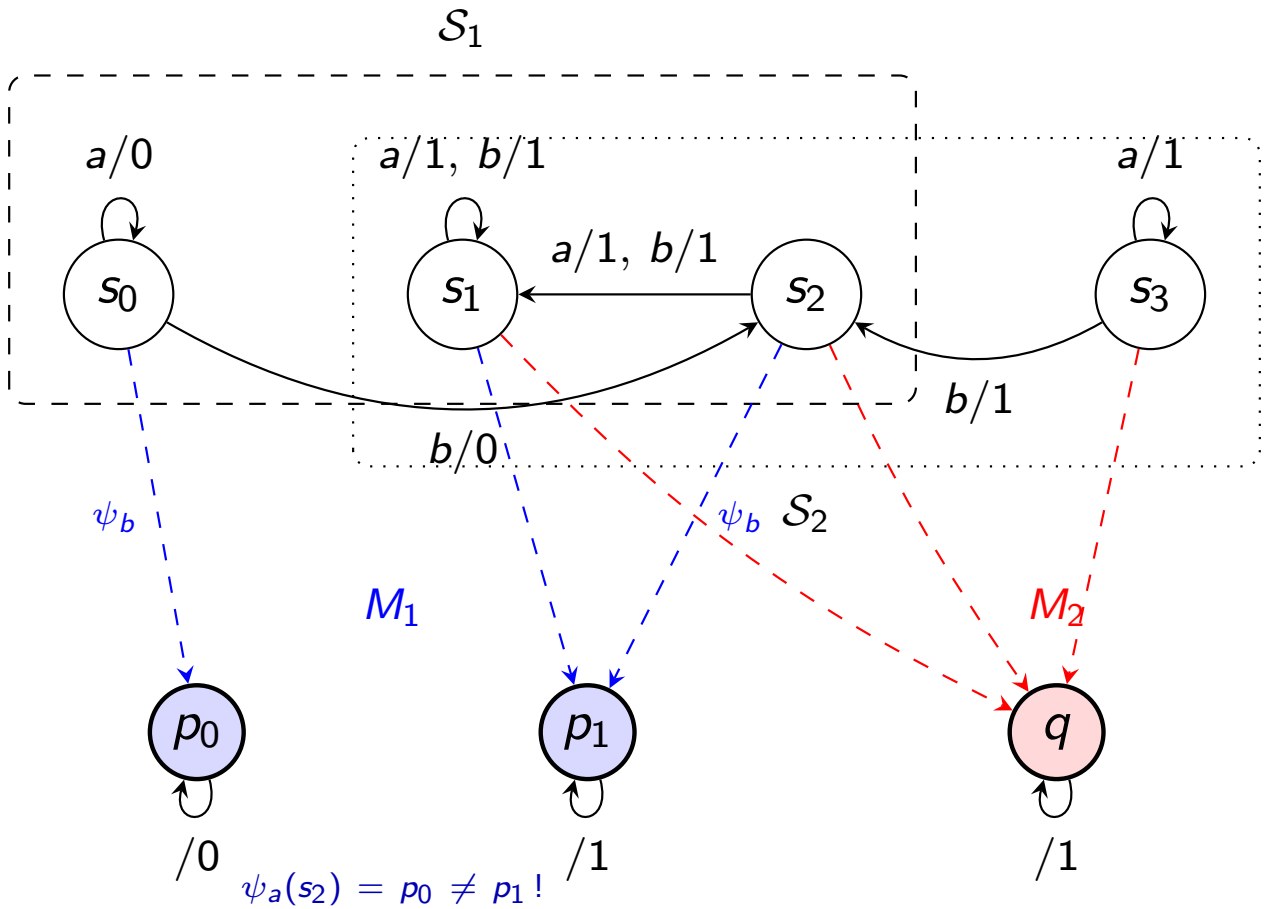
[Rut00] J. J. M. M. Rutten. "Universal coalgebra: a theory of systems". *Theoretical Computer Science*. Modern Algebra 249.1 (2000), pp. 3–80.

[Wan26] Paul Wang. *Separation and Gluing of Explanations on Sites of Dynamical Systems*. 2026. arXiv: 2603.17141 [math.CT].

[WGY18] Gail Weiss, Yoav Goldberg, and Eran Yahav. "Extracting Automata from Recurrent Neural Networks Using Queries and Counterexamples". *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 2018, pp. 5247–5256.

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

Bonus: gluing failure for $\mathcal{F}_j^{\mathrm{beh}}$



Judge: $j_I : \{a, b\} \to \{\bullet\}$, $j_O = \mathrm{id}$. Compatible on $\{s_1, s_2\}$: both give constant output 1.

Bonus: gluing failure — the obstruction

A morphism $\psi : \mathcal{S}_U \to \mathcal{S}'$ has **two state maps**: $\psi_b : S_{U,b} \to S'$ and $\psi_a : S_{U,a} \to S'$.
These can differ, even when $S_{U,b} = S_{U,a}$.

Dynamics square at (s, i):
$\alpha'(\psi_b(s), j_I(i)) = (\psi_a(\alpha_S(s, i)),\ j_O(\alpha_O(s, i)))$.

Since $j_I(a) = j_I(b) = \bullet$, each state gives **two equations with the same LHS**:

- From s_0: $\psi_a(s_0) = \psi_a(s_2)$ (forced by $\alpha(s_0, a) = s_0$, $\alpha(s_0, b) = s_2$)
- From s_3: $\psi_a(s_3) = \psi_a(s_2)$

So $\bar{\psi}_a(s_0) = \bar{\psi}_a(s_2) = \bar{\psi}_a(s_3) =: p$ for any global section.

Bonus: gluing failure — the contradiction

From s_0: the global section must restrict to M_1 on $\mathcal{S}_1$.

$\bar{\psi}_b(s_0)$ has constant-0 behaviour. $\alpha'(\bar{\psi}_b(s_0), \bullet) = (p, 0)$.

$\Rightarrow$ The tail of the constant-0 tree is constant-0:

$\mathrm{Beh}(p) = (0, 0, 0, \ldots)$.

From s_3: the global section must restrict to M_2 on $\mathcal{S}_2$.

$\bar{\psi}_b(s_3)$ has constant-1 behaviour. $\alpha'(\bar{\psi}_b(s_3), \bullet) = (p, 1)$.

$\Rightarrow$ The tail of the constant-1 tree is constant-1:

$\mathrm{Beh}(p) = (1, 1, 1, \ldots)$.

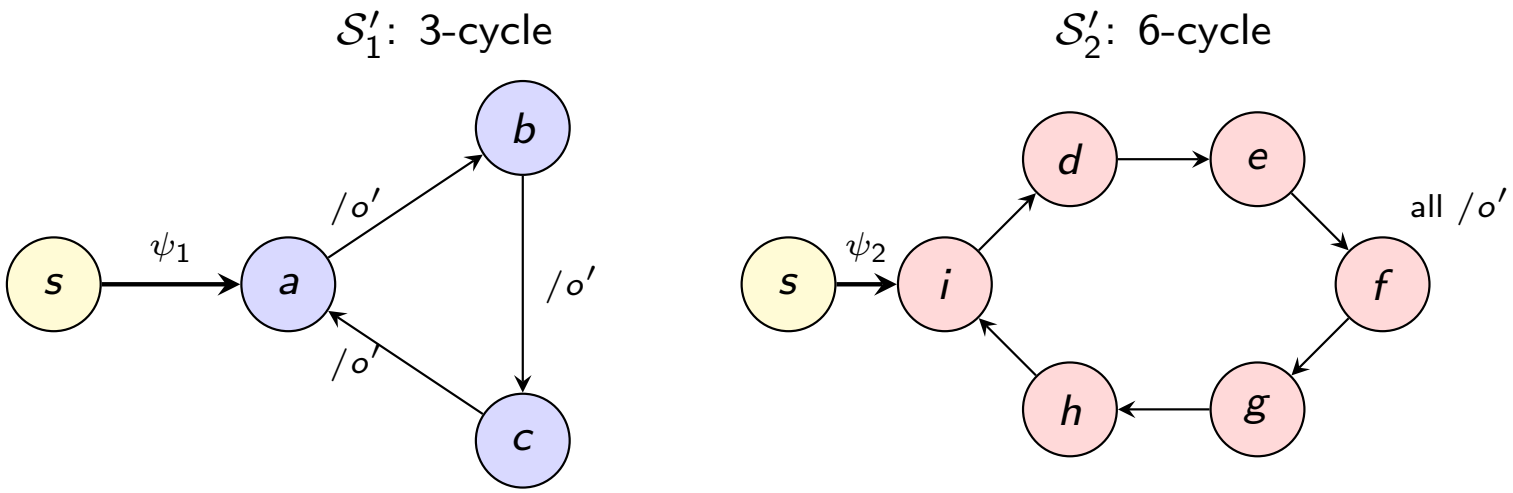
Contradiction

$p = \bar{\psi}_a(s_2)$ cannot simultaneously have behaviours $(0, 0, \ldots)$ and $(1, 1, \ldots)$.

The compatibility condition constrains $\bar{\psi}_b$ on the overlap $\{s_1, s_2\}$, but $\bar{\psi}_a(s_2)$ is only constrained by the dynamics square from states *outside* the overlap.

Bonus: cogerm $\neq$ behavioural equivalence

Two explanations of a trivial system (one state, one input, constant output o'):



Behaviorally equivalent: both produce constant output o' forever.

Not cogerm-equivalent: the 6-cycle has no proper subsystems, so any witness $\mathcal{S}'' \hookrightarrow \mathcal{S}_2'$ has $|\mathcal{S}''| = 6$. No mono from 6 states into the 3-cycle.

Bonus: why this distinction matters

Cogerm equivalence (cf. JNW bisimulation [JNW96]): two explanations share a common subsystem witness. Remembers *internal structure*.

Behavioral equivalence (cf. Rutten [Rut00]): same input-output trees. Only sees *observable predictions*.

Cogerm $\Rightarrow$ behavioural, but **not conversely** (the cycle example).

Consequences for the presheaf hierarchy

- $\mathcal{F}_j^{\mathrm{cogerm}}$: gluing **yes** (adhesivity), separation **no**
- $\mathcal{F}_j^{\mathrm{beh}}$: separation **yes** (pointwise), gluing **no**

For interpretability, we care about *predictions*, not wiring $\Rightarrow$ behavioural equivalence is the natural choice. The cost: we lose gluing.