ICCC'26 TCS & CC
June 29th 2026

From Mechanistic to Compositional Interpretability

Ward Gauderis* · Thomas Dooms*
Steven Homer · Kola Ayonrinde · Geraint Wiggins

VUB ARTIFICIAL INTELLIGENCE RESEARCH GROUP
Universiteit Antwerpen
fwo Flanders AI Research Program

Interpretability and creativity

Ward Gauderis

1. What are compositional interpretations?

Ward Gauderis

2. What makes explanations good?

Thomas Dooms

3. How do we find better explanations?

Ward Gauderis

Payoff and implications

Thomas Dooms

Interpretability and creativity both want to judge the *process*, not only the product.

Post-hoc interpretability

"What information was used to make this particular decision?"

(Jordanous, *Four PPPPerspectives on computational creativity in theory and in practice*. 2016; Sharkey, *Open Problems in Mechanistic Interpretability*. 2025)

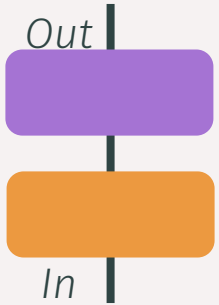
Interpretability and creativity both want to judge the *process*, not only the product.

Post-hoc interpretability

"What information was used to make this particular decision?"

Mechanistic interpretability

"How does the model solve this general class of problems?"



(Jordanous, *Four PPPPerspectives on computational creativity in theory and in practice*. 2016; Sharkey, *Open Problems in Mechanistic Interpretability*. 2025)

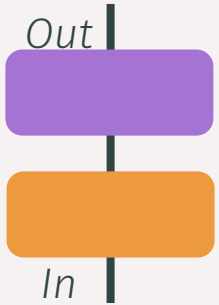
Interpretability and creativity both want to judge the *process*, not only the product.

Post-hoc interpretability

"What information was used to make this particular decision?"

Mechanistic interpretability

"How does the model solve this general class of problems?"



Computational creativity

"To what extent is the behaviour of this system creative?"

Evaluation of Producer, Product, **Process** and Press

- No just-so stories
- Out-of-distribution

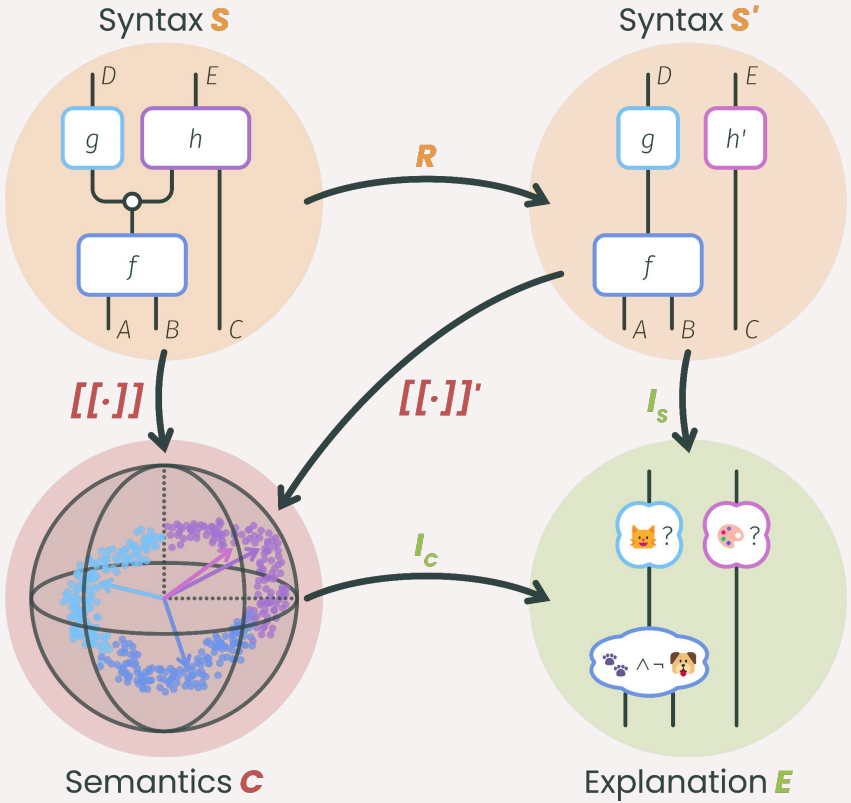
(Jordanous, *Four PPPPerspectives on computational creativity in theory and in practice.* 2016; Sharkey, *Open Problems in Mechanistic Interpretability.* 2025)

Compositional interpretability provides a formal framework to evaluate and improve explanations

Two basic principles, formalised,

- **Compositionality**
  - Category Theory
- **Occam's Razor**
  - Minimum Description Length

cast interpretability as **constrained optimisation**.



(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Interpretability and creativity

Ward Gauderis

1. **What are compositional interpretations?**

Ward Gauderis

2. What makes explanations good?

Thomas Dooms

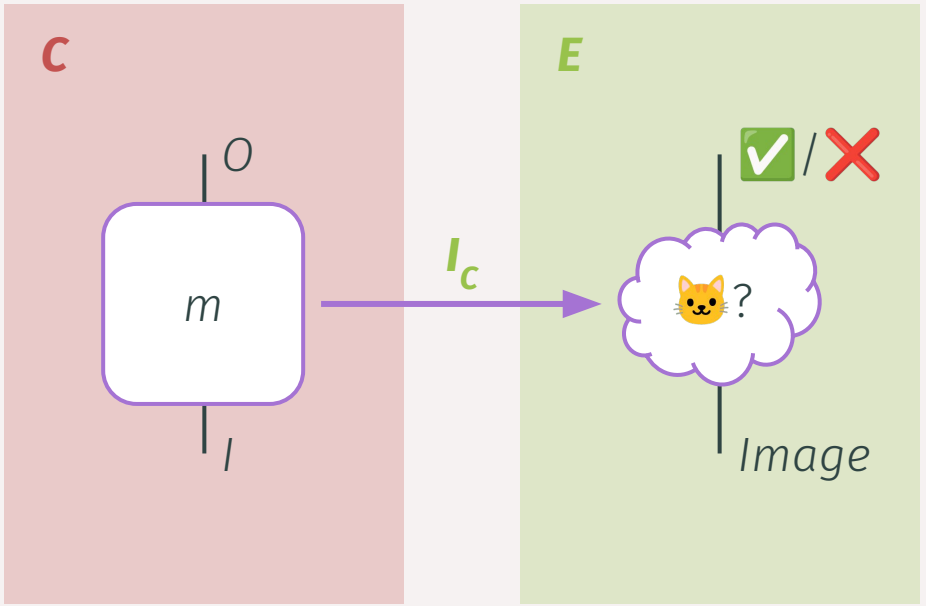
3. How do we find better explanations?

Ward Gauderis

Payoff and implications

Thomas Dooms

Post-hoc interpretations only map input-output behaviour to explanations.

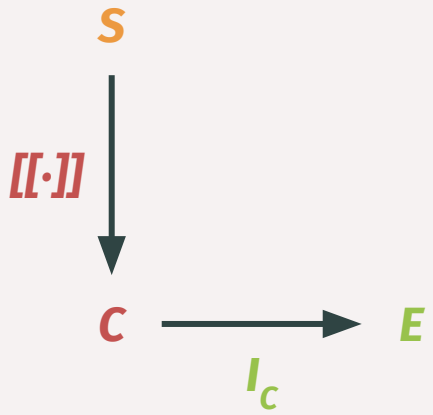


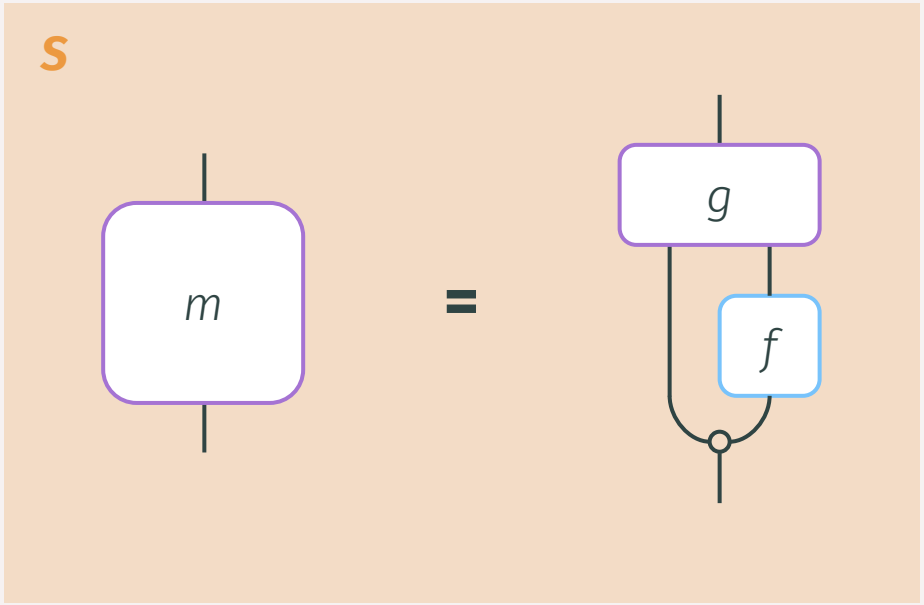
C Semantics
E Explanation

I_C Semantic interpretation

(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Compositional models abstract high-level syntax from low-level semantics.



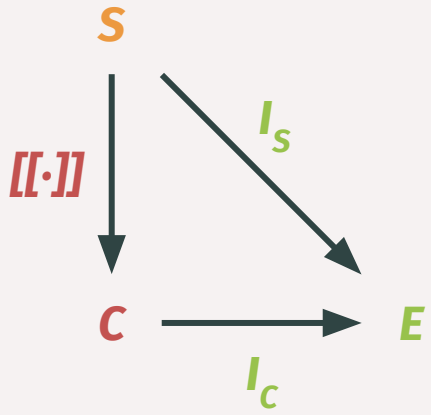


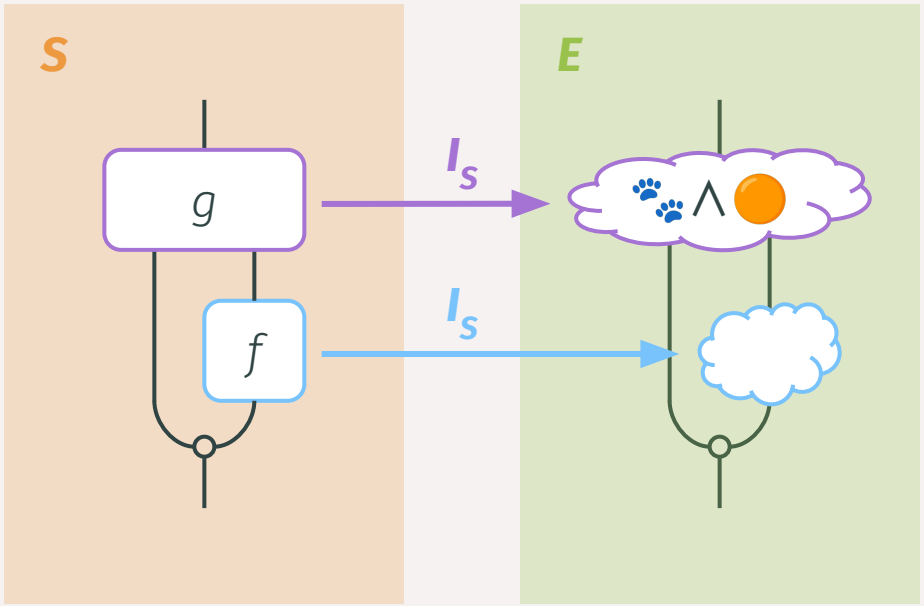
S Syntax
C Semantics
E Explanation

[[·]] Representation
I_C Semantic interpretation

(Tull, *Towards Compositional Interpretability for XAI*. 2024; Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Compositional interpretations mechanistically explain how a model works, step-by-step.





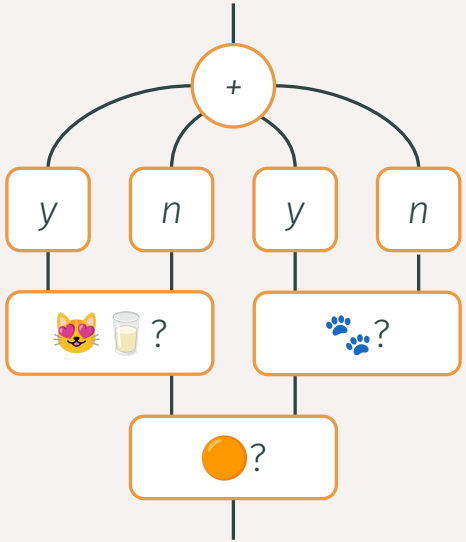
S Syntax
C Semantics
E Explanation

[[·]] Representation
I_C Semantic interpretation
I_S Syntactic interpretation

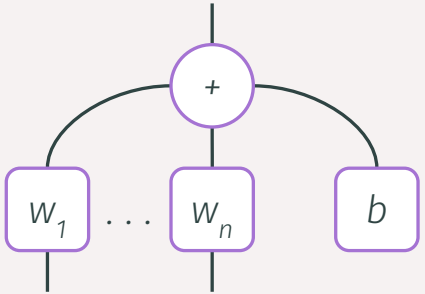
(Tull, *Towards Compositional Interpretability for XAI*. 2024; Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Interpretability methods aim to construct complete compositional interpretations.

Decision tree



Linear regression



Intrinsically interpretable models have compositional interpretations.

(Tull, *Towards Compositional Interpretability for XAI*. 2024; Madsen, *Interpretability Needs a New Paradigm*. 2024)

Interpretability and creativity

Ward Gauderis

1. What are compositional interpretations?

Ward Gauderis

2. **What makes explanations good?**

Thomas Dooms

3. How do we find better explanations?

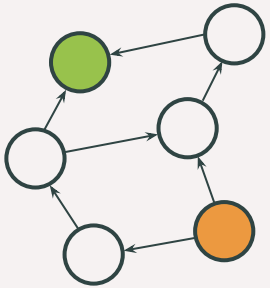
Ward Gauderis

Payoff and implications

Thomas Dooms

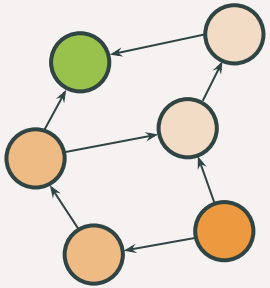
What makes a fastest path algorithm good?

Brute force



exact

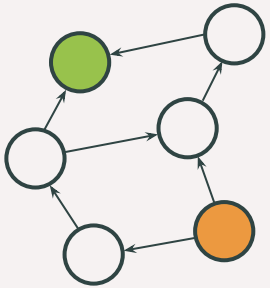
Dijkstra



exact

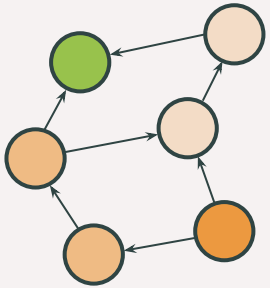
What makes a fastest path algorithm good?

Brute force



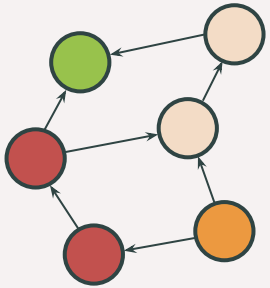
exact

Dijkstra



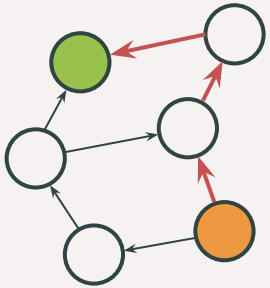
exact

Weighted A*



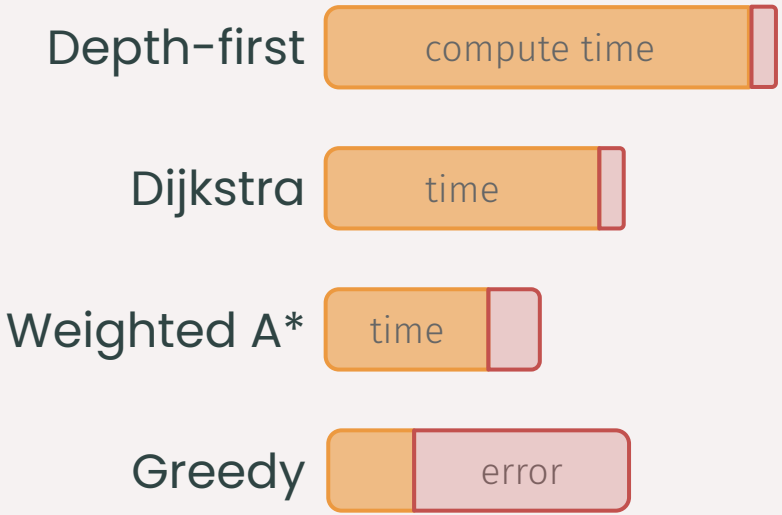
bounded error

Greedy



unbounded error

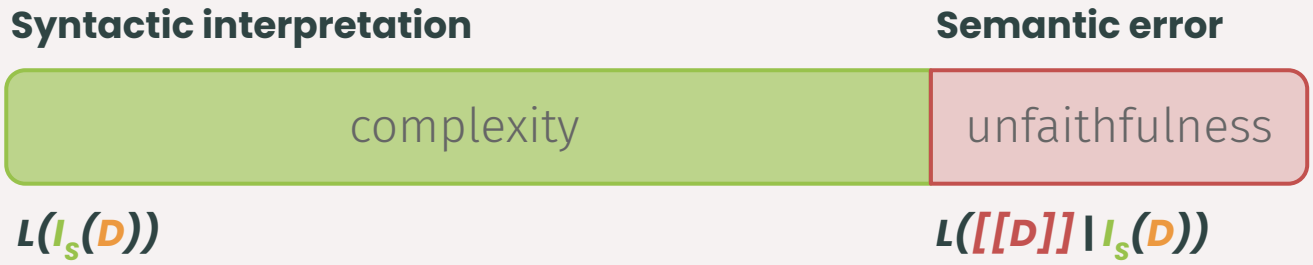
We can rank an algorithm goodness based on time and error.



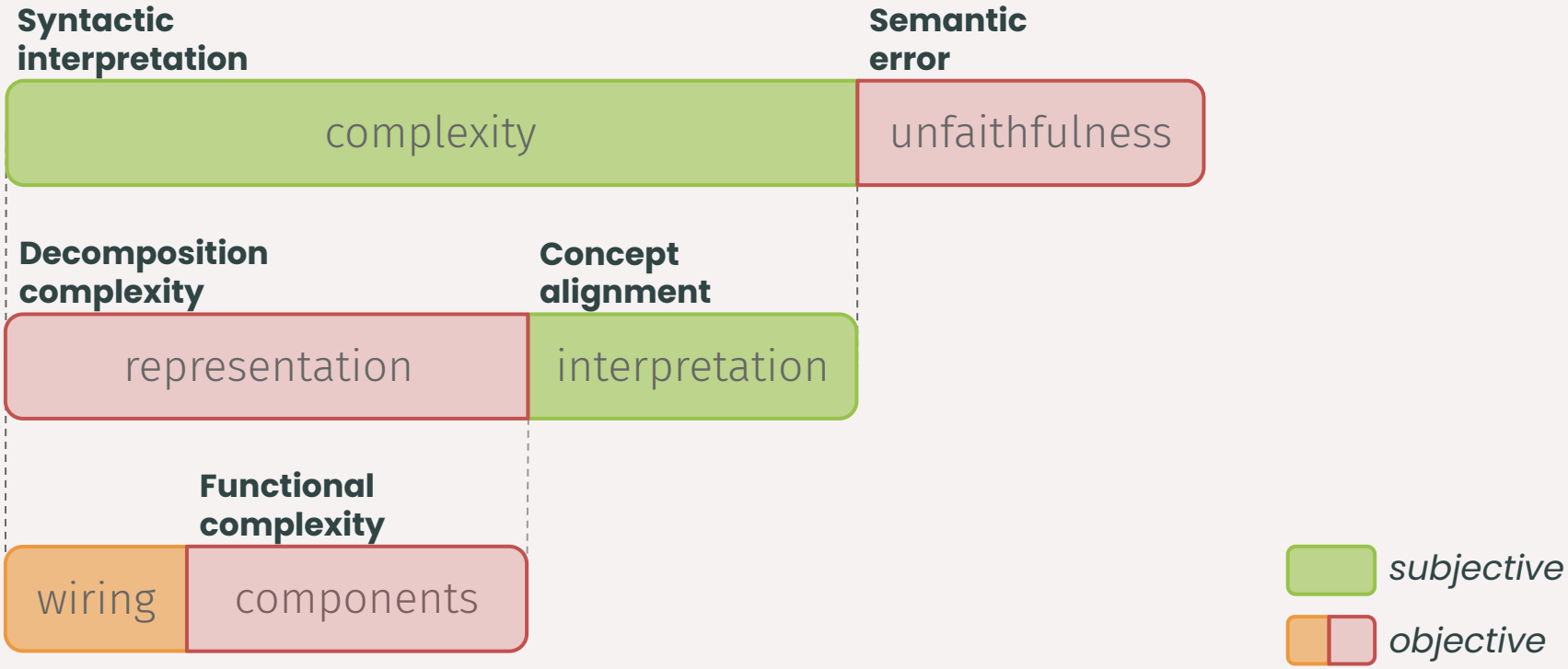
For explanations, we formalise this ranking with the notion of description length.

description length

$$L(D) = L(M) + L(D \mid M)$$



For interpretations, we formalise this ranking with the notion of description length.



Interpretability and creativity

Ward Gauderis

1. What are compositional interpretations?

Ward Gauderis

2. What makes explanations good?

Thomas Dooms

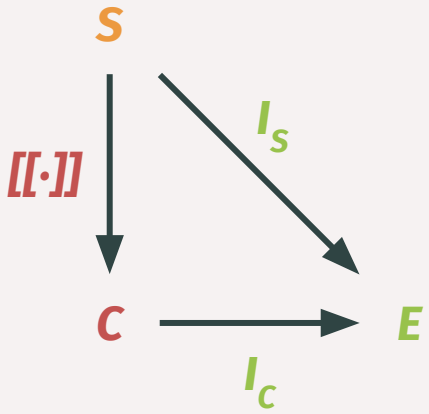
3. **How do we find better explanations?**

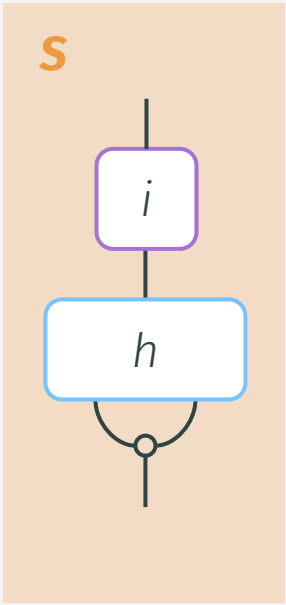
Ward Gauderis

Payoff and implications

Thomas Dooms

Compressive refinements can simplify representations without changing model behaviour.





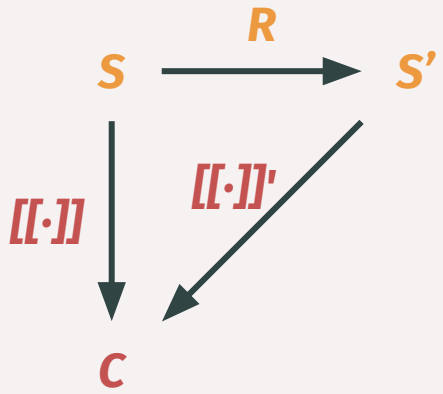
$L([[\cdot]])$

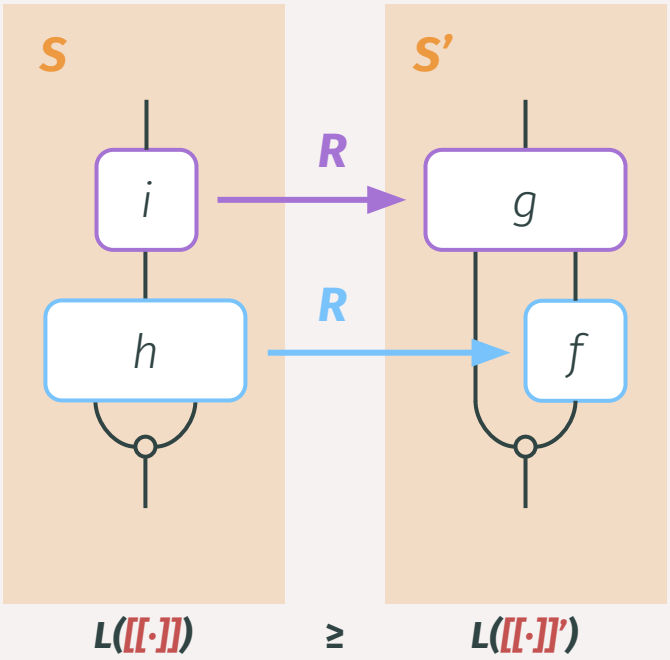
S Syntax
C Semantics
E Explanation

[[·]] Representation
I_C Semantic interpretation
I_S Syntactic interpretation

(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Compressive refinements can simplify representations without changing model behaviour.



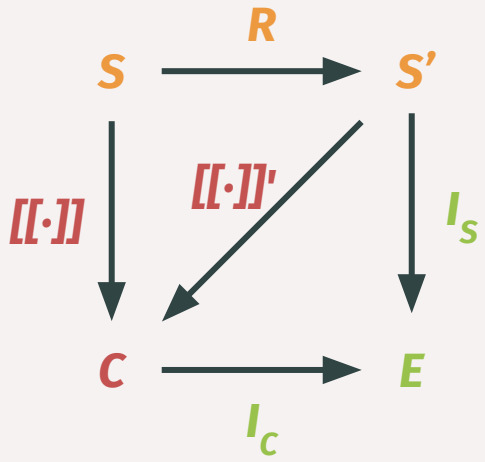


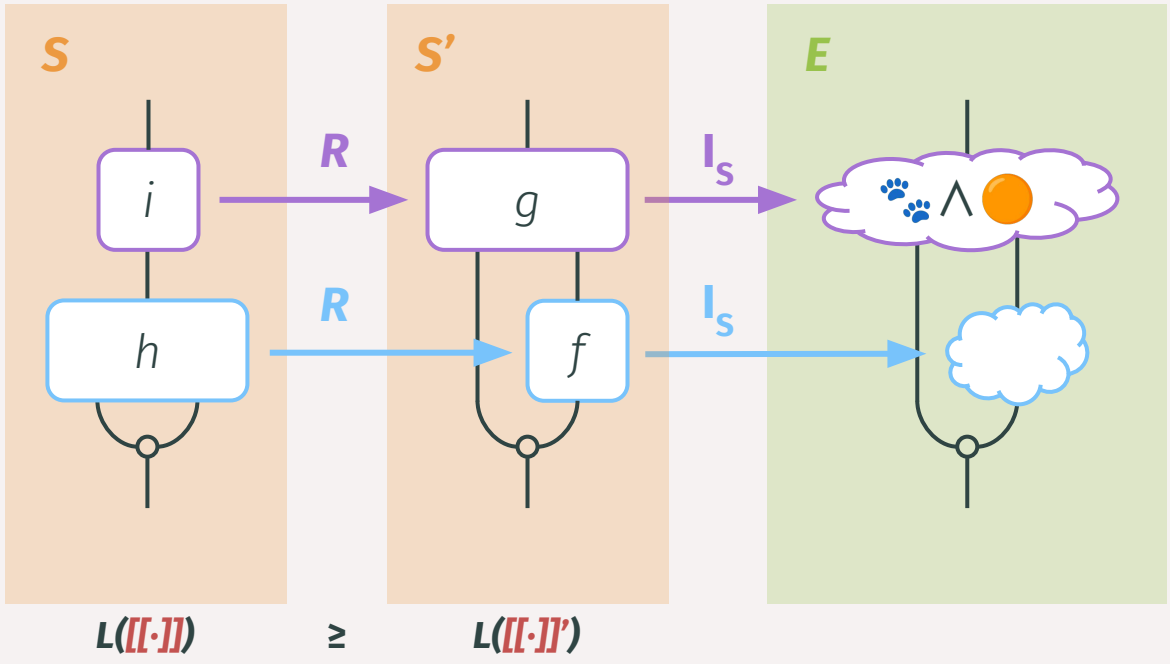
S Syntax
C Semantics
E Explanation

R Syntactic refinement
[[·]] Representation
I_C Semantic interpretation
I_S Syntactic interpretation

(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Compressive refinements can simplify representations without changing model behaviour.



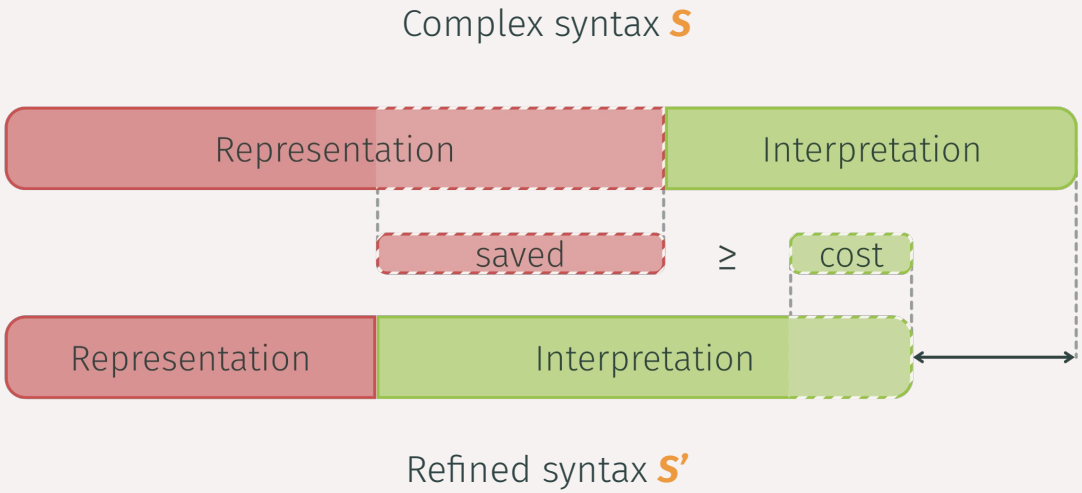


S Syntax
C Semantics
E Explanation

R Syntactic refinement
[[·]] Representation
I_C Semantic interpretation
I_S Syntactic interpretation

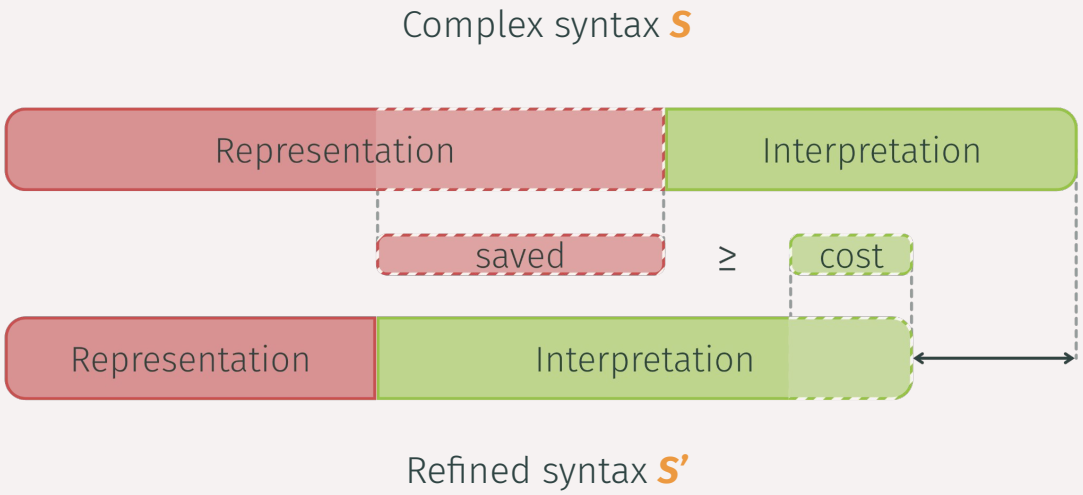
(Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Simpler does not always mean more interpretable… But it does under the *parsimony criterion*.



(Ayorinde, *A Mathematical Philosophy of Explanations in Mechanistic Interpretability*. 2025; Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Simpler does not always mean more interpretable…
But it does under the *parsimony criterion*.



Comonotonic coding suffices.

A coding distribution is a prior.
Encode *explanatory optimism* in it,
and both complexities fall together.

$$L([\![\cdot]\!]) := -\log_2 P([\![\cdot]\!])$$

(Ayorinde, *A Mathematical Philosophy of Explanations in Mechanistic Interpretability*. 2025; Gauderis, Dooms, *From Mechanistic to Compositional Interpretability*. 2026)

Interpretability and creativity

Ward Gauderis

1. What are compositional interpretations?

Ward Gauderis

2. What makes explanations good?

Thomas Dooms

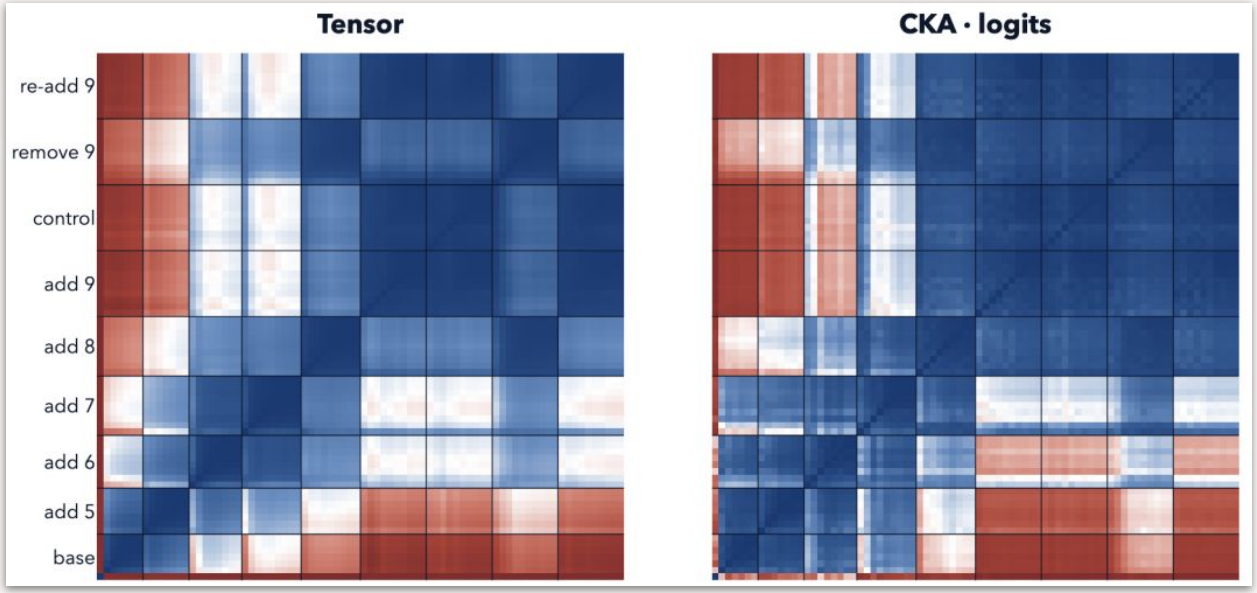
3. How do we find better explanations?

Ward Gauderis

Payoff and implications

Thomas Dooms

Measuring similarity across model training.

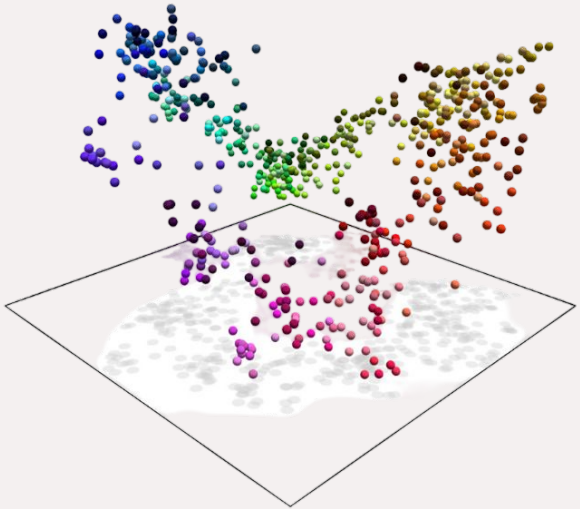


ours

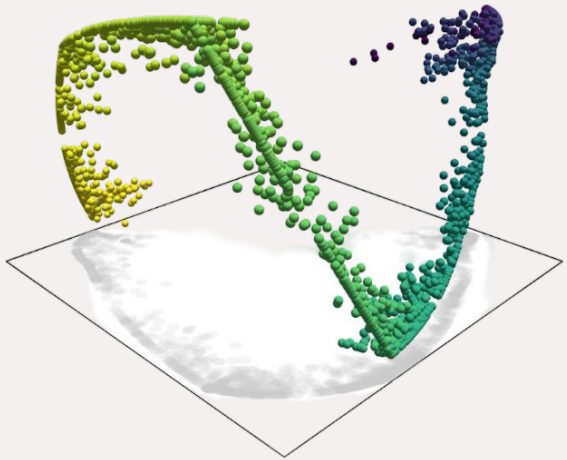
(Nissen et al.; When are two networks the same? tensor similarity for mechanistic interpretability. 2026)

Interpretability aims to find the geometry of concepts inside neural models.

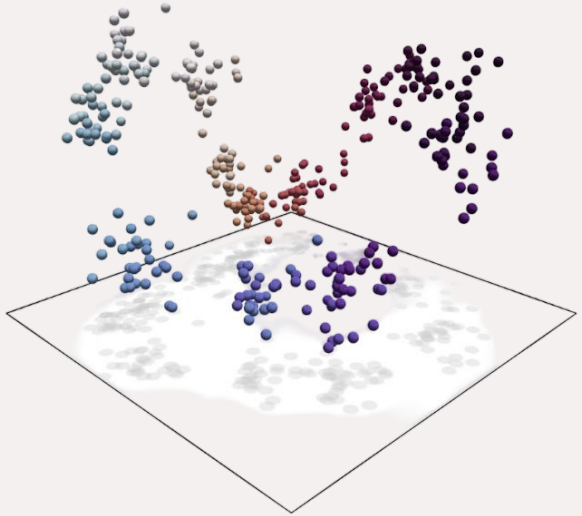
Names for colors



Years of 20th century



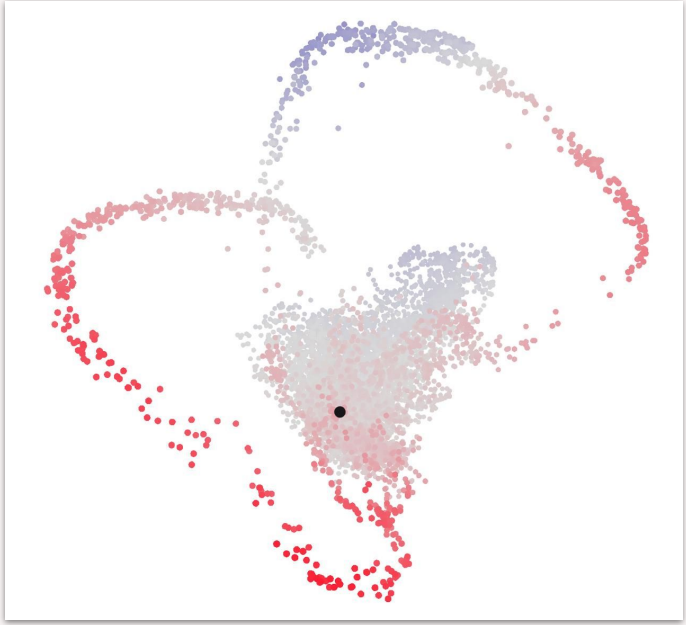
Dates of the year



(Modell et al.; The origins of representation manifolds in LLMs. 2025)

Interpretability aims to find the geometry of concepts inside neural models.

Protein structure

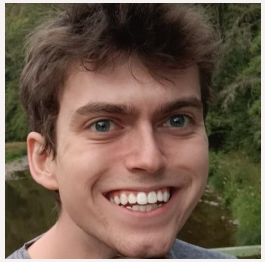


(Internal experiments)

Want to know more or reach out?

Ward Gauderis

Geraint Wiggins



Thomas Dooms

Jose Oramas

VUB ARTIFICIAL INTELLIGENCE RESEARCH GROUP

Universiteit Antwerpen