# From Mechanistic to Compositional Interpretability

Extended abstract of arXiv:2605.08934 – Workshop on Theoretical CS and Computational Creativity, ICCC'26
Ward Gauderis*[1], Thomas Dooms*[2], Steven T. Homer[2], Kola Ayonrinde[3], Geraint A. Wiggins[1]
[1]Vrije Universiteit Brussel [2]Independent [3]UK AI Security Institute

Computational creativity has evaluated the product, formalising an artefact's novelty and value [5], more than the process behind it [2]. Judging the process calls not for a story that reproduces the output but for an explanation faithful to the original computation, a need it shares with mechanistic interpretability. We introduce *compositional interpretability*, a category-theoretic framework grounded in compositionality and minimum description length (MDL) that makes an explanation's quality measurable and optimisable; agnostic to what produces an artefact, it applies to creative systems as well as neural networks.

**Motivation.** Mechanistic interpretability aims to reverse-engineer a model into human-understandable components through decomposition, description, and validation [6], yet without a formal framework, such explanations cannot be verified, compared, or composed. The decomposition fixes the structure that the other stages rest on: if its parts do not recombine into the whole, locally correct descriptions stay globally uninformative. We thus seek interpretations that are *compositional*, *faithful* (matching internal computation), and *concise* (having low description length for a human interpreter).

**Compositional explanations.** A compositional model [7] separates a system's **syntax**, its abstract wiring (e.g. neural layers), from its **semantics**, the functions those parts implement. Formally, a string diagram $D$ in a syntactic category $S$ is given semantics by a representation functor $[\![\cdot]\!] : S \to C$ in a semantic category $C$. An interpretation consists of two groundings into human-understandable terms $H$: a structural one $I_S : S \to H$ and a behavioural one $I_C : C \to H$, required to commute as $I_S = I_C \circ [\![\cdot]\!]$. This forces structure and behaviour to agree, ruling out just-so stories that fit behaviour alone.

**Evaluating explanations.** At a meta-level, MDL treats the model's semantics as data and the explanation as the code compressing it, so explanation quality splits into *faithfulness*, the information needed to correct the explanation back to the true semantics, and *complexity*, its description length. Faithfulness ranges from *behavioural* (reproducing outputs) to *compositional* (reproducing mechanisms); complexity splits into an objective *representation* cost of the compositional model and a subjective *interpretation* cost of aligning components with human concepts.

**Optimising explanations.** Following Kowalski's *Algorithm = Logic + Control* [3], we fix the semantics and vary only the syntax, optimising for intelligibility rather than efficiency. Grounded this way, faithfulness and complexity are both computable, casting interpretability as a rate-distortion problem: a semantics-preserving *compressive refinement* $R : S \to S'$ minimises representation complexity under a faithfulness threshold. A *parsimony criterion* gives the condition under which reducing representation complexity also lowers total description length, guaranteeing more concise, human-aligned explanations. Prominent mechanistic methods can be cast as subclasses of refinement, clarifying why their compressibility heuristics tend to align with human interpretability.

**Relevance to computational creativity.** For creativity, three consequences follow. First, a creative process should be judged by compositional, not merely behavioural, faithfulness, reconstructing the generative computation rather than reproducing the artefact. Second, because creativity depends on generalisation beyond the inspiring set, this faithfulness must hold out of distribution. Third, conciseness is not value [4]: our parsimony criterion identifies when compressing a system's representation also improves its explanation. Bounded-observer measures like epiplexity [1] raise the question of whether what counts as creative or interpretable may depend on the observer's own computational capacity.

**References** **[1]** M. Finzi et al. *From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence.* arXiv:2601.03220 [cs]. 2026. **[2]** A. Jordanous. "Four PPPPerspectives on computational creativity in theory and in practice". *Connection Science* 28.2 (2016), 194–216. **[3]** R. Kowalski. "Algorithm = Logic + Control". *Commun. ACM* 22.7 (1979), 424–436. **[4]** T. Mondol and D. G. Brown. "Computational Creativity and Aesthetics with Algorithmic Information Theory". *Entropy* 23.12 (2021). **[5]** G. Ritchie. "Some Empirical Criteria for Attributing Creativity to a Computer Program". *Minds Mach.* 17.1 (2007), 67–99. **[6]** L. Sharkey et al. *Open Problems in Mechanistic Interpretability.* 2025. **[7]** S. Tull et al. *Towards Compositional Interpretability for XAI.* 2024.