Binding Before Revision: Symbolic Structure and the Admissibility Boundary in Neural Representations

Flyxion
Independent Researcher

September 2026

Abstract

Neural networks frequently develop approximately symbolic internal representations, according to a recent paper by McCoy, Soulos, Linzen, and Smolensky, which supports the claim not by decoding structure from a network but by replacing a network's representations with explicit symbolic approximations while preserving most of its behavior. Their method, DISCOVER, models a representation as a linearly transformed tensor product representation and tests the approximation causally, through substitution, intervention, and generalization to unseen bindings. This essay summarizes the method and its evidentiary structure, assesses the strength and limits of the argument, and then situates the result within a framework distinguishing operations within an admissibility space from operations on that space. The central claim is that DISCOVER's compositional bindings, however systematic, remain transformations of commitments inside a fixed domain of admissibility; they do not by themselves constitute the distinct operation of revising that domain.

1 The DISCOVER Method and Its Evidence

McCoy et al. propose that a neural representation can often be modeled as a linearly transformed tensor product representation (TPR), in which a filler is bound to a role and these bindings are summed and passed through an affine map:

$$E_{\text{TPR}} = W\left(\sum_i f_i \otimes r_i\right) + b.$$

In a sentence, for instance, a particular word might serve as filler and its grammatical function as role. Their method, DISCOVER ("DISsecting COmpositionality in VEctor Representations"), learns the filler vectors, role vectors, and affine transformation that best approximate a target network's internal activations, and then substitutes the reconstructed vectors back into the network in place of the originals. If the untouched decoder continues to produce correct output, the approximation is functionally adequate rather than merely correlated with recoverable information.

The experiments proceed from synthetic list-manipulation networks, where bidirectional positional roles approximate representations well across architectures including multilayer perceptrons, recurrent networks, bottleneck Transformers, and ordinary Transformers, to seven large language models: Gemma-3-27B, GPT-2-XL, GPT-OSS-20B, Pythia-12B, Qwen3-14B, OLMo-2-13B, and Llama-3.1-8B. The most extensive case study examines GPT-OSS across arithmetic, syllogistic logic, Python execution, passivization, tense reinflection, and question formation. Replacing internal representations with DISCOVER approximations produces an average performance gap of only

2.36 percentage points from the original model. Surgical edits to the role–filler structure produce predicted behavioral changes: altering a number changes an arithmetic answer, moving a code element between lists changes execution, and changing a modifier's grammatical role causes the model to behave as though the modifier had appeared elsewhere in the sentence. Across thirty-one kinds of intervention, mean intervention accuracy reaches 0.903.

A further test asks whether DISCOVER can handle role–filler combinations withheld from training. Successful generalization in most conditions matters because an unconstrained TPR could otherwise disguise an atomic lookup table with no genuine compositional structure. White-box controls sharpen the point: DISCOVER generalizes to unseen bindings when the target network is constructed to use systematic TPR structure, and fails to generalize when the target instead uses independent embeddings for every role–filler pair. The authors read the overall pattern through Smolensky's "limitivism": neural networks neither eliminate symbols nor implement perfectly discrete symbol systems, but approach symbolic organization while retaining approximation, noise, gradience, and statistical flexibility.

2 Assessment of the Argument

The paper's strength lies in combining three kinds of evidence that are usually kept separate: representational approximation, causal intervention, and out-of-distribution generalization. A conventional probe can show only that role information is recoverable from a vector; it cannot show that the network relies on that organization to produce its output. The replacement and intervention experiments make the causal claim substantially stronger than a probing result would.

The design also anticipates the most obvious objection to any symbolic reconstruction: that sufficiently large filler and role embeddings could encode every binding independently, so that successful in-distribution reconstruction alone would say nothing about genuine composition. The held-out-binding experiments, the regularization used to fit the model, and the white-box comparisons provide substantial evidence against this possibility.

The most striking single result concerns period representations in one experiment, where a decoder trained on real GPT-OSS vectors reconstructs sentences more accurately from DISCOVER's approximations than from the original vectors, with accuracy rising from 0.71 to 0.96 in the reported layer. This suggests that DISCOVER is extracting and idealizing a noisy structural component the decoder already partially uses, which supports the intermediate reading of "approximately symbolic" better than either the claim that the representation is merely vectorial or the claim that it implements a conventional discrete symbol system.

Several qualifications remain. DISCOVER begins from researcher-specified fillers and candidate role schemes, so the method is better described as hypothesis testing than unrestricted discovery: the investigator supplies a prospective ontology, and the method determines whether the network's vectors are compatible with it. A successful fit does not establish that the chosen role scheme is unique, minimal, or native to the model. Relatedly, linearly transformed TPRs span a broad family of vector-symbolic architectures, so the results support multiplicative binding followed by additive aggregation as a description without determining whether the network implements tensor products as a computational mechanism, or which specific vector-symbolic architecture is the closest fit. The tasks are also deliberately controlled and often templatic, which is appropriate for causal analysis but leaves open whether comparably clean decompositions exist in unrestricted conversation, ambiguous language, long-context reasoning, metaphor, conflicting instructions, or tasks whose relevant roles are not known in advance; the most extensive cross-domain analysis, moreover, concentrates on a single model, GPT-OSS, even though the period-representation results span

seven.

The word "emergent" also needs care. The structures in question were not explicitly built into the architecture, but the models were trained on data produced by highly symbolic human practices: language, mathematics, logic, and programming. The paper provides strong evidence that approximately symbolic structure emerges through learning on such data; it does not demonstrate that symbolic organization would emerge independently of symbolically organized training data. Finally, an intervention accuracy of 0.903 is high but incomplete: the unsuccessful cases could reflect noise, an imperfect role scheme, information distributed outside the edited representation, or genuinely non-symbolic computation, and these residuals are theoretically significant rather than mere experimental slack. The paper is also a first arXiv version, dated August 30, 2026, with only a partial codebase released pending employer approval, so independent reproduction is not yet established.

3 Binding Within a Fixed Admissibility Space

The result fits cleanly into a distinction between operations within an admissibility space and operations on that space. A DISCOVER representation presupposes a vocabulary of admissible fillers, roles, and bindings; given that vocabulary, it describes how a particular structure is realized inside vector space. This is an account of inference or transformation within a fixed domain of admissibility, denoted Ω. It is not by itself an account of how a system creates a genuinely new role, refuses an existing category, excludes a class of representations, or revises the space of admissible structures.

In this notation, DISCOVER mostly concerns transformations of commitments inside W_Ω, the space of admissible states given Ω. Even a role-changing intervention replaces one admitted binding with another admitted binding; it does not ordinarily perform

$$A_R^-(\Omega) = \Omega \setminus R,$$

the removal of a role from the admissible vocabulary itself. This reinforces the claim that zero weight, suppression, replacement, and refusal are distinct operations. A filler can be removed from one represented structure without its type being removed from the system's admissible vocabulary; vector surgery changes the instantiated expression without necessarily revising the grammar under which expressions can be formed.

Indeed, changing a role is not yet changing the inventory of roles. DISCOVER can move a filler from the role of subject adjective to that of object adjective, but both roles and the operation of reassignment remain antecedently available. This is structural recombination, not structural legislation. Revision begins only when the system can alter which roles, bindings, or transformations count as admissible.

The paper also refines the shorthand that reweighting is not refusal. Neural computation need not be an unstructured reweighting of undifferentiated features: its continuous vectors can carry compositional role–filler organization, as DISCOVER shows. But discovering that richer structure does not collapse the difference between modifying a state and modifying its domain. Symbolic binding explains how elements occupy positions within a representational space; admissibility explains which positions, elements, and transformations are permitted to exist in the first place.

The compositional-generalization results also map onto a distinction between representation and continuation. A network can possess a compositional internal representation without reliably extending that composition to unseen cases. Possessing a structured encoding does not guarantee an admissible continuation operator that preserves the structure under novelty; the representational

resource can exist while the transition policy fails to use it consistently. The same point bears on questions of stated principles and behavioral stability under pressure. A model may encode a principle, role, or distinction in a systematic way without being governed by it across contexts. DISCOVER demonstrates causal reliance within bounded tasks; it does not demonstrate stable commitment under adversarial prompting, social pressure, conflicting objectives, or distribution shift. The familiar problem persists in a new register: an agent's ability to represent a rule does not establish that the rule constrains its later conduct.

The method itself illustrates a broader warrant question. DISCOVER does not autonomously reveal a single true symbolic organization; investigators propose an admissible family of interpretations and test which members preserve behavior under substitution and intervention. The evidence warrants the claim that a TPR-style structure is sufficient and causally relevant to the network's behavior on the tasks tested; it does not warrant the stronger claim that this structure is the model's uniquely correct ontology. The distinction is between an effective reconstruction and an exhaustive explanation, and the paper is careful, on its own terms, not to claim the latter.

4 Conclusion

Recent mechanistic evidence complicates the description of neural networks as systems that merely reweight undifferentiated features. McCoy et al. show that continuous representations can approximate compositional structures in which fillers are systematically bound to roles, and that this organization is causally implicated in behavior rather than merely decodable from it. Yet the result strengthens, rather than dissolves, the distinction between inference within a space and revision of the space itself. Replacing one filler–role binding with another modifies a representation under an existing grammar; it does not remove the role, filler, or transformation from the system's domain of admissibility. Symbolic structure within vector space supplies a richer account of W_{Ω}, the space of states defined over an admissibility domain, but not an operator that transforms $\boldsymbol{\Omega}$ itself. The paper offers strong evidence that neural representations can be systematically symbolic without being discretely symbolic; it does not show that such systems possess stable principles, autonomous ontological revision, or constitutive refusal, and its silence on those points marks where the admissibility framework begins.

References

[1] R. Thomas McCoy, Paul Soulos, Tal Linzen, and Paul Smolensky, "The Emergent Symbolic Structure of Artificial Neural Networks," arXiv preprint arXiv:2608.29530, version 1, 30 August 2026. https://arxiv.org/abs/2608.29530.