Behavioural vs. Representational Systematicity in End-to-End Models: An Opinionated Survey

Ivan Vegner[1,4]* **Sydelle de Souza**[1,4]* **Valentin Forch**[2]
Martha Lewis[3]‡ **Leonidas A.A. Doumas**[4]‡

[1]School of Informatics, University of Edinburgh
[2]Chemnitz University of Technology
[3]Institute of Logic, Language, and Computation, University of Amsterdam
[4]School of Philosophy, Psychology & Language Sciences, University of Edinburgh

Abstract

A core aspect of compositionality, systematicity is a desirable property in ML models as it enables strong generalization to novel contexts. This has led to numerous studies proposing benchmarks to assess systematic generalization, as well as models and training regimes designed to enhance it. Many of these efforts are framed as addressing the challenge posed by Fodor and Pylyshyn. However, while they argue for systematicity of *representations*, existing benchmarks and models primarily focus on the systematicity of *behaviour*. We emphasize the crucial nature of this distinction. Furthermore, building on Hadley's (1994) taxonomy of systematic generalization, we analyze the extent to which behavioural systematicity is tested by key benchmarks in the literature across language and vision. Finally, we highlight ways of assessing systematicity of representations in ML models as practiced in the field of mechanistic interpretability.

1 Introduction

A core feature of human cognition is the ability to understand and generate novel combinations of known concepts in systematic ways. To illustrate, understanding (and producing) the utterance "*the octopus ate the fish*" implies the ability to also understand (and produce) "*the fish ate the octopus*". This *systematicity*, often considered fundamental to the concept of compositionality, has become an increasingly important evaluation criterion for artificial intelligence systems. Recent years have seen a proliferation of papers proposing both benchmarks to test for systematicity (e.g., Lake and Baroni, 2018; Hupkes et al., 2020; Kim and Linzen, 2020; Wu et al., 2023; Kim et al., 2023; Okawa et al., 2023) as well as new model architectures designed to exhibit systematic behaviour (e.g., Locatello et al., 2020; Didolkar et al., 2021; Soulos et al., 2024; Assouel et al., 2024). Many of these works frame their contributions as addressing the challenge[1] posed by Fodor and Pylyshyn (1988, F&P): to explain how systematic language competencies can emerge without structured mental representations. However, there is a crucial distinction that is often overlooked. F&P specifically argued for the necessity of systematic representations in cognitive systems, while modern benchmarks and evaluations typically focus solely on testing for the presence of systematic behaviour. This conflation has led to misunderstandings and conflicting claims about the systematic generalization capabilities of artificial neural networks. This is further compounded by black-box nature of our models.

This paper aims to disentangle these issues by separating behavioural and representational systematicity. We argue that meaningful progress toward human-like systematic generalization requires appropriately-scoped behavioural claims backed up by rigorous mechanistic interpretability. We develop our argument in three parts. First, we discuss the historical development of the concept of systematicity. Drawing on lessons from psychology, we highlight how the focus has gradually shifted from representations to behaviour, as well as the impact of this shift. Second, we analyse key benchmarks in the language[2] and vision literature, assessing the type of systematicity they evaluate using Hadley's (1994) framework of weak, quasi-, and strong systematicity. Finally, we explore current approaches to evaluating systematicity of representations in end-to-end trained Transformer models through the lens of mechanistic interpretability. Our review concludes with a set of recommendations for evaluating claims of systematicity in machine learning (ML) systems. It must be noted that works measuring compositionality of representations using

*Joint first authors ‡ Joint senior authors

[1]Although F&P do not explicitly propose a challenge, it is referred to as such in Fodor and McLaughlin (1990).

[2]Specifically, at the syntactic and semantic levels.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31842–31856
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

custom architectures (e.g., Andreas (2019); McCoy et al. (2020); Li and McClelland (2022)), while relevant, are beyond the scope of this paper.

Related Work Compositionality has received a great deal of attention in ML. However, although one of the most frequently invoked definitions of compositionality is derived from the human capacity for systematicity, it has largely been sidelined as a secondary concern.

McCurdy et al. (2024) survey the NLP research community, asking them to rate their agreement with a definition of "compositional behaviour" (CB) and a set of statements regarding the investigation thereof. They define CB as a model's tendency to produce a correct output for input I given that it produces correct output for human-defined component parts of I. They further decouple their definition from any notion of learning, making their definition of CB agnostic to the dataset(s) from which it is learned. We believe that systematicity is indeterminable under this definition. Absent a clear definition of the component parts and the contexts in which they appear during training, it is impossible to judge a model's ability to systematically recombine them. This impairs even behavioural evaluation of a model's generalization capabilities.

Russin et al. (2024) provide a comprehensive historical overview of compositionality as an object of inquiry. They also review current trends in endowing ML architectures with compositional mechanisms. However, their central question of "can contemporary neural networks replicate the behavioral signatures of compositionality" (p. 11) also focuses on behaviour rather than representation. They only briefly discuss systematicity and do not engage with Hadley's (1994) distinctions between levels of systematicity, which is central to our analysis. Nevertheless, they do call for mechanistic interpretability of model representations to reinforce conclusions in behavioural tests of compositionality.

There are some reviews which discuss representational systematicity, albeit in different terms. Wattenberg and Viégas (2024) survey key mechanisms for enabling neural networks to represent relationships between features—a long-standing question in cognitive science, known as the binding problem (see Feldman, 2013, for an overview). They propose a number of directions for future work with respect to interpreting how trained models may be representing (or approximating) relational binding. Greff et al. (2020) argue that contemporary neural networks lack human-level compositional generalization abilities precisely because they are unable to solve the binding problem. The authors provide a comprehensive theoretical review of the connection between the binding problem and compositionality. They present an extensive survey of mechanisms which may implement binding and the properties which such mechanisms should possess. Conversely, Pavlick (2023) argues that existing LLMs can already encode symbols and undertake symbolic processing. Despite their differences, these papers all underscore the importance of mechanistic interpretability in establishing the presence of systematic representations in neural models.

2 Compositionality via Systematicity

Often attributed to Frege (1892), the principle of compositionality is frequently stated as: "the meaning of a complex expression is a function of the meanings of its parts and of the way they are syntactically combined" (Partee, 2004). It is essential in formal semantics, serves as a foundational element in most logical formal languages[3], and its influence extends across modern linguistic theories (Halvorsen and Ladusaw, 1979; Gazdar et al., 1985; Steedman, 2019; Joshi, 2005). In technical terms, this principle can be understood as a mathematical relationship—syntax and semantics function as algebras, with meaning assignment acting as a homomorphism that maps syntax to semantics (Janssen and Partee, 1997).

It is crucial to note that the principle of compositionality is "extremely theory-dependent" (Partee, 2004). Without establishing a theory of syntax, semantics, and the interaction between them, the principle is severely underspecified and can be made trivial if not completely vacuous (Westerståhl, 1998; Kazmi and Pelletier, 1998). We echo Janssen and Partee (1997): "the principle [of compositionality] should not be considered an empirically verifiable restriction, but a methodological principle that describes how a system for syntax and semantics should be designed." We now turn to F&P's definition of compositionality, which puts systematicity in the driver's seat.

[3]Although there are some exceptions, see for example, Hintikka (1979).

2.1 Fodor and Pylyshyn's Compositionality

In order to establish a syntax and semantics of the language of thought, Fodor and Pylyshyn (1988, F&P) build their definition of compositionality from the more basic notion of systematicity. They define systematicity as an entailment between cognitive capacities: a human capable of understanding the statement aRb, where a, b are objects and R is a relation, is necessarily able to understand bRa. For example, for the vast majority of language users, understanding the sentence "*John loves Mary*" necessarily implies the ability to understand and produce the sentence "*Mary loves John*", and vice versa. F&P argue that this capacity is evidence for the existence of structure and structure-sensitive operations in the human brain. They assert that this property arises because such sentences are composed from the same constituents: "insofar as a language is systematic, a lexical item must make approximately the same semantic contribution to each expression in which it occurs." F&P admit relaxations of this principle in language, such as context-sensitive word definitions, syncategorematic compositions such as "good [X]" and idioms. However, they maintain that for most language users to be able to make systematic linguistic inferences, human cognitive representations must be systematic too.

Importantly, F&P position their definition of compositionality—consisting of systematicity and the related properties of productivity and inferential coherence—as a property of the representational system (specifically of human mental representations). For F&P, composition is a syntactic (structural) operation which produces a semantic result. By complement, the semantics of individual linguistic units are amenable to performing structural operations over them. They assert that in representations without such structure, human-like compositionality is incredibly unlikely to emerge. Indeed, F&P's challenge is to explain how systematicity can arise in a representational system which does not posit structural operations over symbols (Fodor and McLaughlin, 1990).

2.2 Hadley's Three Levels of Systematicity

Based on the work of F&P, Hadley (1994) outlines three progressive levels of systematicity in language learning.

Weak systematicity is defined as the ability to process sentences that use familiar words in new combinations, but only if those words have appeared in identical syntactic positions before elsewhere in the training data. For instance, if a system is trained on sentences like "*John loves Mary*," "*Jim loves John*," and "*Mary loves Jim*," it should be able to process "*Mary loves John*" as a test case.

A system possesses *quasi-systematicity* if it has weak systematicity under recursion. Quasi-systematic models can process novel sentences that contain embedded clauses, provided both the main sentence and the embedded clause are structurally isomorphic to various sentences in the training corpus. For any successfully processed complex sentence, the training data must include a simple sentence demonstrating the same word in the same syntactic position in which it appears in the embedded clause.

To exhibit *strong systematicity*, the system must be able to process novel sentences, including those with embedded clauses, even when words appear in syntactic positions that they did not occupy in the training data in any simple or embedded clause. Hadley states that strong systematicity is closest to human systematic generalization, as the system can integrate its understanding of syntax and semantics to process words in entirely novel structures.

2.3 Systematicity and Productivity

Closely connected to systematicity is the notion of linguistic productivity, defined as the ability to make "infinite use of finite means" (Chomsky, 1965, quoting Humboldt (1836)). To disambiguate the two, many have redefined productivity to refer specifically to a capacity for understanding and producing unbounded inputs and outputs (e.g., Fodor and Pylyshyn, 1988; Hupkes et al., 2020), or "length generalization" in modern parlance.

We conjecture that Hadley's weak systematicity will yield no productivity, because for a sentence of an unseen length there will necessarily be at least one word in the sentence in a novel syntactic position. A quasi-systematic model could exhibit productivity for some, but not all, sentences given its ability to process some recursive clauses. Finally, strong systematicity yields full productivity.

3 From Representation to Behaviour

The core of the distinction between representational and behavioural systematicity lies in the classical distinction between competence and performance (Chomsky, 1965, p. 2). Representational systematicity is a competence, a cognitive mechanism which ideally manifests itself in performance, i.e., systematic behaviour, except as mitigated by other factors such as compute limitations, noise, errors, etc. Below, we lay out the implications of this distinction for ML research.

3.1 Operationalisation

F&P defined systematicity as a relation between cognitive competencies. Hadley reformulated their definition in terms of the properties that must be true of learning systems claiming to possess systematicity. In other words, he put forth a framework for the *operationalisation* of systematicity.

Operationalisation is a common practice in fields like psychology, wherein unobservable properties of black-box systems must be reasoned about via proxy measures. For example, to study an unobservable property like memory function, a psychologist may operationally define memory in terms of recall from a list. Operationalisation, though, is not without its potential pitfalls. For any operational definition A of a concept B, demonstration of A is at best only evidence for the presence of B, and the strength of the evidence depends on the strength of the connection between concept and operationalisation. Psychologists have long grappled with these issues, leading to the development of subfields like measurement theory and psychometrics (e.g., Luce, 1996; Stevens, 1946).

Consider the use of the relational match to sample task in comparative cognition (e.g., Premack and Woodruff, 1978). In this task, the participant is presented with a pair of sample items (e.g., two red shapes) and asked to select one of two additional pairs of items that best corresponds to the sample (e.g., two blue shapes vs a red shape and a blue shape). The task aimed to test the capacity to reason about a relational concept like same/different, and was historically used as a demonstration that certain animals had or did not have this capacity (e.g., Thompson et al., 1997; Fagot and Parron, 2010). However, Mike Young and colleagues (Fagot et al., 2001; Young and Wasserman, 1997, 2001) demonstrated that certain animals solved the task simply by detecting and responding to the entropy in sets of items, without recourse to relational reasoning.

The converse problem exists as well. The field of developmental psychology has long dealt with the issue of performance and competence with children, for whom failure to exhibit a behaviour does not necessarily imply the lack of capacity to do so in any circumstance (such as due to distraction or disinterest). In the literature on object permanence, for example, there has been a long debate about if and when children actually possess the capacity (e.g., Piaget, 2013; Baillargeon and DeVos, 1991). A resolution one way or the other would circumscribe the scope of what a child is able to do and understand at a given age. However, this has proven elusive due to the difficulty of operationalising object permanence appropriately.

3.2 Implications for Machine Learning

There are parallels to be drawn here with research in ML. The literature is rife with failures of operationalisation, in which a model's performance on a particular task poorly predicts performance on other, even very similar tasks (see Geirhos et al., 2020, for an overview of "shortcut learning"). Thus, modellers must ensure that their tasks can only be solved using the competence which they intend to test.

However, demonstrating perfect operationalisation is very difficult. Consequently, when evaluating generalization capabilities, researchers require insights into the underlying representations that drive the behavioural outcomes on their task of choice. This *mechanistic interpretability* enables researchers to assess the validity of their operationalisation and predict model performance across tasks that draw upon the same competence, even when behavioural tests prove imperfect.

According to F&P, systematic generalization requires systematicity of internal representations (*representational systematicity*), not merely evidence of systematicity in the model's performance on any number of tasks (*behavioural systematicity*). The relationship between valid operationalisation, behavioural evidence, and mechanistic interpretability creates three analytical cases, each with different implications for our understanding of model capabilities:

Case 1: Systematic behaviour without valid operationalisation. Valid operationalisation requires demonstrating that tasks actually demand specific levels of behavioural systematicity, rather than weaker levels or no systematicity at all. This demands principled control over dataset syntax and semantics. If this is not established, success or failure on the behavioural task no longer constitutes evidence of the existence or absence of representational systematicity. Models may rely on various non-systematic strategies to achieve performance, even when datasets intuitively appear to require systematicity. Sections 4 and 5.2 examine several imprecise operationalisations of systematicity and demonstrate how this inhibits drawing meaningful conclusions about the systematic generalization capabilities of evaluated models.

Case 2: Systematic behaviour under valid operationalisation. When models demonstrate systematic behaviour under conditions of valid operationalisation, F&P assert that they must possess systematic representations. If one gives credence to this position, supported by Hadley and other theorists of structured cognition, this case will be evidence of systematic internal representations. However, definitively proving this requires mechanistic interpretability of the model representations underlying systematic behaviour. Demonstrating that a model exhibits systematic behaviour without possessing systematic representations would effectively overcome Fodor and Pylyshyn's challenge, representing a significant theoretical breakthrough.

Case 3: No systematic behaviour under valid operationalisation. This case also permits two competing interpretations. The model may fail to display systematic behaviour because it does not have systematic representations. Alternatively, the model may possess systematic representations, but be unable to use them to behave systematically.

The second possibility is analogous to the child who cannot, or does not want to, demonstrate object permanence in a particular context. This possibility becomes more worrisome as ML systems become more complex and modular, with components performing one subtask pipelined with components solving another. Even if one module has representations that support systematicity, the system will be unable to demonstrate systematicity on the task if a non-systematic module intervenes before the output layer. One example of models ending up in this case (partially) as a result of dataset construction is illustrated in Wu et al. (2023), which we discuss below. As in the case above, distinguishing between these two interpretations requires mechanistic interpretability of the model's representations. Both are consistent with F&P's challenge.

4 Behavioural Systematicity

We propose that Hadley's taxonomy provides a robust framework for operationalising systematic generalization. It clearly defines the generalization capabilities of each level of systematicity, and establishes the requirements for testing it in a particular train/test data split. Below, we use it to examine representative datasets that test behavioural systematicity in end-to-end trained models. Rather than attempting an exhaustive review of this rapidly evolving field, we focus on key examples that demonstrate the critical importance of evaluating the level of systematicity required by a particular dataset. We hope that researchers will find this analysis useful and in turn apply it to their models and datasets.

4.1 Linguistic Tasks

SCAN, NACS Lake and Baroni (2018) introduce SCAN, a synthetic dataset for testing compositionality in text. It consists of all 20,910 command phrases generated from a non-recursive phrase-structure grammar of 13 words. Each command unambiguously maps onto a sequence of actions (for example, "jump opposite left and walk thrice" maps to "LTURN LTURN JUMP WALK WALK WALK"). It aims to test compositionality via a judicious set of data splits: Split 1 is an iid (independent, identically distributed) split with up to 99% of the training datapoints withheld randomly. Split 2 tests a form of productivity by testing the network on phrases which map to longer action (output) sequences than any that occur in training. Split 3 asks the network to parse compositions with a held-out word that it had only ever seen in a primitive context before (e.g., parsing "run twice and jump" having only seen "run" as a primitive, but having also seen "look twice and jump" before).

We suspect that Split 1 of SCAN tests Hadley's definition of weak systematicity. Lake and Baroni (2018) state that with even 98% of the training set for Split 1 withheld, every command in the test set that does not have a conjunction also appears in the training set an average of 8 times. Given that conjunction-less commands must also appear in conjunctions, it is likely that even 2% of Split 1 exhibits all possible commands in all possible positions. Split 2 tests productivity and thus requires at least quasi-systematicity[4]. Finally, Split 3 aims to test strong systematicity sans recursion. However,

[4]The fact that the SCAN grammar is non-recursive does not mean the model cannot rely on recursion internally.

Bastings et al. (2018) argue that good performance on SCAN can be achieved with simple modelling tricks that require neither structured representations nor systematicity, demonstrating the difficulty of inferring anything about a model's representations based on the model's behaviour on the task.

PCFG SET Hupkes et al. (2020) propose the PCFG SET dataset, an artificial translation task similar to SCAN. In this task, input sequences—generated by a probabilistic context-free grammar (PCFG) and corresponding to string edit operations—must be "translated" into output sequences representing the result of the sequence of operations. The authors propose several splits which aim to test various aspects of (their definition of) compositional generalization[5]:

The **Systematicity split** tests the model's ability to produce the correct output for a sequence of operations *a b*, if it has not seen this (sub)sequence in the training dataset, but has seen *a* and *b* in other contexts. This split appears to test Hadley's weak systematicity. Although, unlike in SCAN, the PCFG SET grammar is recursive and therefore the dataset is non-exhaustive, it is likely that most operations will individually have been seen in most of the positions in which they will occur in the test dataset. This is because the parameters for constraining the length and recursive depth of the generated input-output pairs are the same for the train and test sets.

The **Productivity split** involves training the model on sequences containing up to 8 operations, and testing it on sequences containing 9+ operations. As argued in Section 2.3, this split tests at least quasi-systematicity, and may contain some cases requiring strong systematic generalization.

COGS, ReCOGS, SLOG Proposed by Kim and Linzen (2020) and extended by Wu et al. (2023) and Li et al. (2023a), the COGS family of datasets examines systematicity in a linguistic semantic parsing context. Directly specifying the meaning of a representation is challenging; instead, semantic parsing involves mapping sentences to logical forms (LFs). It does so under the assumption that the LFs can then be mapped to meanings more directly (Wu et al., 2023). The task involves parsing multiple semantically-equivalent expressions into a (more) canonical semantic representation concerned with binding the semantic roles of events and actions to particular entities. Here, we discuss the original COGS dataset (Kim and Linzen, 2020), but our arguments apply equally to its variants.

The dataset operationalises semantic parsing as a sequence-to-sequence problem. The model's task consists of predicting the LFs (as strings) for a large set of natural-language sentences. Systematicity is evaluated by testing the model on a generalization set, constructed by systematically altering the sentences in the training set. These manipulation include:

- **Novel pairings of primitives and syntactic roles** (e.g., Subject $\rightarrow$ Object and vice versa)
- **Alternations of verb structure** (e.g., "Charlie blessed Emma" $\rightarrow$ "Lina was blessed")
- **Inferences based on verb class**, i.e., the semantic roles of a verb (e.g., "The cobra helped the dog" $\rightarrow$ "The cobra froze")
- **Deeper recursion** than seen in training
- **Object modifiers to subject modifiers** (e.g., "Noah ate the cake on the plate" $\rightarrow$ "The cake on the table burned")

To the extent that the first three cases, termed "lexical generalization" in the paper, evaluate novel recombinations of familiar items and structures, and the fourth and fifth, termed "structural generalization", test the ability to process novel structures, all test cases require strong systematicity.

The changes proposed by ReCOGS (Wu et al., 2023) illustrate the importance of the distinction between behavioural and representational systematicity. Compared to COGS, ReCOGS (1) removes irrelevant output tokens, (2) isolates the effects of length vs. depth generalization and (3) mitigates spurious correlations between syntactic position and variable indexing. None of these transformations alter the meanings of the LFs and, therefore, the nature of the capabilities which COGS purports to test; nevertheless, ReCOGS evokes much better performance on structural generalization from the same models evaluated in Kim and Linzen (2020). Thus, it is hard to draw any conclusions as to the true systematic generalization capabilities of any model evaluated on COGS, as any model's success

[5] The authors also propose splits testing the useful properties of Substitutivity, Localism and Overgeneralization, which are out of the scope of the current work. Dankers et al. (2022) extends this framework to look for evidence of these factors in a natural-language machine translation dataset, and highlighting the difficulty of rigorously evaluating compositionality in real-world data.

or failure on COGS may be due to spurious and unintended effects of dataset construction, rather than the model's capability for systematic representation (see Csordás et al., 2021). We note that SLOG (Li et al., 2023a) incorporates the proposals of ReCOGS and adds 17 new systematic generalization cases, providing a compelling behavioural benchmark for strong systematicity.

4.2 Visual Tasks

In contrast to linguistic tasks, compositionality and systematicity in vision have only recently re-emerged as a focus of research. While perception is undoubtedly compositional—complex representations are made up of simpler percepts and relations between them—the nature of compositionality in perception is contested (see Lande, 2024). Furthermore, properties such as systematicity and productivity have not been clearly defined for visual perception and must not necessarily mirror their linguistic definitions (Cummins, 1996; Lande, 2024; Camp, 2007). Lacking a strong theory of the syntax of images, it is unclear how Hadley's levels of systematicity apply to visual concepts (see Cavanagh, 2021; Lande, 2021; Quilty-Dunn et al., 2023). Multiple datasets address these challenges by testing compositionality in computer vision through three distinct approaches.

The first approach involves the **disentanglement of generative variables**. Here, the aim is to test whether a model can represent stimuli as a conjunction of several properties or variables which are conditionally-independent given the stimulus (Schott et al., 2022). Specifically, models are trained on compositionally constructed inputs to either directly predict the variables of the underlying generative model (Schott et al., 2022), classify composites of real-world images (Liao et al., 2023), or, given an input image I, produce an image I' which only differs with respect to a minimal change in the generative variables (e.g., exchange of colors between two objects while preserving their shape, size, and position) (Kim et al., 2023). These datasets can only test weak systematicity, as they (a) lack a hierarchical structure, precluding them from incorporating recursion, which is necessary for quasi-systematicity, and (b) do not require the models to make use of concepts in novel contexts (e.g., colour always relates to a shape's interior, never its outline).

The second approach involves **vision-based abstract reasoning tasks** designed to evaluate conceptual understanding rather than mere perceptual feature composition. Models must first infer the underlying invariants across images, and use this information to produce an appropriate response—e.g., identifying rule violations (Zerroug et al., 2022), choosing the right image from a candidate set (Zhang et al., 2019; Odouard and Mitchell, 2022), or generating rule-based solutions (Chollet, 2019; Moskvichev et al., 2023; Assouel et al., 2022). The RAVEN domain (Zhang et al., 2019) has no hierarchical structure, although Zerroug et al. (2022) make use of hierarchical rule compositions. The open-ended ARC domain (Chollet, 2019) allows for complex task structures as well as novel applications of concepts, but there is no systematic description of its construction process. Thus, ARC may constitute a test for strong compositionality but cannot be judged on an objective basis. Derivative works (Moskvichev et al., 2023; Assouel et al., 2022) give clearer task descriptions but still lack a rigorous control over the presentation of concepts in novel contexts, also making them unsuitable for testing strong systematicity.

A third approach examines **behavioural systematicity in vision-language models** without committing to a theory of compositionality in vision, leveraging linguistic theories of systematicity to test their text component. However, they do not exhaustively manipulate the syntax of the image caption. Instead, they approximate it by placing novel atoms and compounds in various syntactic slots in the caption and test whether the model's behaviour is sensitive to such manipulations (e.g., Johnson et al., 2017; Thrush et al., 2022; Ma et al., 2023; Lewis et al., 2024). These approaches either test for weak systematicity (e.g., Johnson et al., 2017; Lewis et al., 2024) as they use indivisible objects and do not always change the contexts in which concepts are used, or can only be applied to pre-trained models (e.g., Thrush et al., 2022; Ma et al., 2023), making it impossible to judge the model's true generalization capabilities given unknown training exposure[6].

4.3 Evaluating Learning Trajectories

Systematicity also relates to the speed of knowledge acquisition—a systematic learner should require training data that scales linearly with the number of concepts, whereas for a non-systematic learner this scaling should be exponential, assum-

[6]For an exemplary discussion of the pitfalls of constructing a dataset testing only for systematicity, see Diwan et al. (2022).

ing that the underlying grammar allows full permutations of concepts. A compositional representation should also allow faster skill acquisition as learning a new task will not require a reconfiguration of the encoding space. Okawa et al. (2023) examine convergence rates of diffusion models, finding that the number of training steps required to learn new concepts indeed scales exponentially with Hamming distance in conceptual space. Thomm et al. (2024) also find a super-linear scaling in an algorithmic task requiring Transformer models to map input strings to the outputs of compositional algorithms.

5 Representational Systematicity

To recap, investigating a model's representations allows us to reason about the model's performance on behavioural tests of systematicity, even in light of imperfect operationalisation[7]. Mechanistic interpretability provides evidence towards compositional, if not explicitly systematic, representations. This includes how predictions are built through a model (nostalgebraist, 2020); how information can be routed, copied, and deleted through a network (Elhage et al., 2021); how a kind of compressed sensing can be used to allow models to represent efficiently (Elhage et al., 2022); and the tension between representational compression versus the flexibility of full compositionality (Olah, 2023). In examining trained models for systematic representations, most work takes the approach of "concept discovery", i.e., looking for the presence and use of a specific concept in a model's activations (Sharkey et al., 2025).

5.1 Evidence for

Tenney et al. (2019a,b) uses linear probing to identify the presence of linguistic concepts that might be used in systematic representations. Having received much attention in recent years, this approach has been applied to the discovery of circuits responsible for algorithmic computation (Bricken et al., 2023; Huben et al., 2024; Park et al., 2024).

Li et al. (2023b) and Nanda et al. (2023) use probing to argue that a Transformer trained on the game Othello learns a 'world model'—a systematic representation of the rules and dynamics of the game that is used to plan the next move[8].

Todd et al. (2024) and Merullo et al. (2024) go further, extracting components (specifically, 'function vectors') of a trained network that can be recombined in other contexts to produce predictable output (cf. Opiełka et al., 2025).

Finally, Feng and Steinhardt (2024) and Feng et al. (2025) discover 'binding vectors' that sufficiently large LLMs use to bind entities. These can be used to encode propositions in linguistic tasks. Whilst their experiment does not test for systematicity (propositions are subject-verb-object, but verbs are such that subject and object cannot be swapped), it does propose a possible mechanism for systematicity in LLMs.

5.2 Evidence Against

Much of the evidence against systematic representations can be summed up in a critique of probing, both linear and non-linear. Specifically, the ability to decode a given property from a model's representations does not guarantee that the model causally relies on this property (Belinkov, 2022; Sharkey et al., 2025). Indeed, the property may be decoded from the representations even if it is uncorrelated to task performance or distributed in the input data as random noise (Ravichander et al., 2021).

Vafa et al. (2024) suggest that world models learned by Transformers on several tasks are "far less coherent than they appear". For example, they train a Transformer on a dataset of taxi journeys in Manhattan, requiring it to output a valid sequence of turns from an origin to a destination. They demonstrate that the Transformer does not learn a consistent world model despite excellent next-token prediction performance. The model learns a number of fictitious and impossible road connections; consequently, when the model is made to detour from its chosen route, the proportion of valid routes generated decreases dramatically.

Furthermore, Aljaafari et al. (2025) show that models do not represent words and phrases as systematic compositions of tokens. Instead, these representations are spread across attention heads and layers. Finally, even when (weakly) systematic representations can be detected, they are not always used. Kobayashi et al. (2024) train a vanilla Transformer model on a task that involves systematic recombination of latent variables for a regression task. The model learns to generalize on in-distribution

[7]Although there is a wide range of neurosymbolic approaches that explicitly bake in systematic representations (e.g., Mao et al., 2019; Coecke et al., 2010; Sen et al., 2022; Badreddine et al., 2022; Doumas et al., 2022), we only discuss to representational systematicity in end-to-end trained models.

[8]However, cf. jylin04 et al. (2024)

tasks, and the vector of underlying latent variables can be decoded from its activations. However, the model does not use this knowledge when faced with OOD tasks, and therefore fails. The Transformer is only able to systematically generalize to OOD tasks when augmented with an additional hypernetwork that explicitly takes the decoded latent vector as input.

Overall, there is mixed evidence for the use of systematic representations in Transformer models. While systematic representations can at times be extracted or elicited from the model, they also frequently resort to non-systematic tricks and shortcuts to achieve performance. This may be due to the conflict between efficient data encoding and the development of fully systematic representations (Olah, 2023). Furthermore, even when models do learn apparently systematic representations, it is difficult to ascertain to what extent they use them on OOD tasks.

6 Takeaways for Machine Learning

Compositionality is rapidly emerging as a key empirical factor of strong, robust and interpretable generalization behaviour in ML systems. However, the study of compositionality draws on a rich theoretical foundation spanning diverse approaches to syntax and semantics. As a result, two definitions of compositionality rooted in different frameworks may have little in common when applied to modelling and prediction. As the volume of research on compositionality in ML systems increases, it is crucial to maintain clear definitions of the phenomena in question. In this work, we have focused on Fodor and Pylyshyn definition, but other definitions are certainly possible. We urge researchers to engage with existing literature and situate their definition in the context of previous work, enabling the field to make fair comparisons amongst the multitude of datasets, methods and models.

While many researchers reference F&P's seminal work, they frequently overlook the fact that F&P define compositionality through systematicity, specifically systematicity of representations rather than behaviour. This is a crucial distinction, as behavioural systematicity does not necessitate the presence of systematic representations. A model might achieve strong performance on systematicity benchmarks through mechanisms that do not involve systematic internal representations, such as memorization or task-specific heuristics. There is evidence that this is already the case—Sun et al. (2023) show that different compositional generalization datasets rank the same modelling approaches differently while purporting to test the same capability. To address this issue, researchers should be explicit about their focus on behavioural systematicity, representational systematicity, or both. Those claiming to address F&P's challenge must demonstrate the presence or absence of systematic representations, while those working on behavioural measures should justify why their operationalisation of systematicity is valid.

When examining systematicity through Hadley's framework of weak, quasi-, and strong systematicity, we find that many current benchmarks fall short of testing strong systematicity—the level that most closely approximates human systematic generalization capabilities. Many datasets inadvertently test only for weak or quasi-systematicity, limiting our ability to assess whether artificial systems can achieve human-like systematic processing. Future benchmark development should explicitly target strong systematicity, with careful consideration of the theoretical requirements for such evaluation. Li et al. (2023a) is an example of such a benchmark. However, strong systematicity is difficult to test for in large-scale pre-trained models, as we are blind to what they see during training. Consequently, claims of strong behavioural systematicity may need to be limited to custom trained models.

Above all, the emergence of mechanistic interpretability techniques offers promising approaches for investigating representational systematicity. Recent work on function vectors and binding mechanisms suggests that some neural networks may indeed have mechanisms required for systematicity, though often in unexpected ways. However, many studies conflate the discovery of interpretable components with proof of systematic processing—the mere presence of decomposable representations does not guarantee they are used systematically, nor that the discovered mechanisms are causally responsible for the model's systematic behaviour.

In sum, it is representations, not task performance, that drive generalization. Until we develop principled methods for controlling training distributions and validating representational mechanisms, the quest for strongly systematic learners will remain stifled by the opacity of existing models. Despite the challenges, we believe that rigorous assessments of representational systematicity are well worth the effort.

7 Limitations

This paper discusses systematicity, an important facet of compositionality, and how it is currently tested across ML literature. Limitations of the paper include the following:

- We concentrate on clarifying the definition of one facet of compositionality, i.e. systematicity. In future work, more aspects of compositionality should be treated in the same manner.
- We have discussed widely used benchmarks in the literature to highlight the points we are making about common practices in ML, rather than an exhaustive summary of all literature.
- The benchmarks we have discussed tend to be somewhat based in English, and our analysis has focussed on monolingual English models. Multilingual models may have different generalization abilities.
- Our current analysis primarily assumes phrase structure grammar frameworks, but different theoretical perspectives on syntax may uncover additional dimensions of systematicity that warrant further investigation. Therefore, a full treatment of systematicity under alternative syntactic theories such as tree-adjoining grammars (Joshi, 2005), construction grammar (Goldberg, 1995), or combinatory categorial grammar (Steedman, 2019) represents an important direction for future work.

8 Ethics

As a survey paper, there is limited ethical impact of the form of bias in datasets or environmental impact of training. As mentioned above, we do concentrate on benchmarks and tasks phrased in English, which may have an impact on the conclusions we draw. However, overall, we feel that our emphasis on thorough testing of claims of systematicity, and call for behavioural systematicity to be backed up by representational systematicity, should have a positive ethical effect in pushing for the development of more reliable and predictable models.

Acknowledgements

IV, SdS, ML, and LD gratefully acknowledge travel and collaboration funds (RIS International Collaboration Grant) from Research Innovation Scotland. SdS, VF, and ML gratefully acknowledge the opportunity to build this collaboration at the SFI Working Group: Representations in Minds and Artificial Systems at the Santa Fe Institute. IV and SdS acknowledge that this work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

References

Nura Aljaafari, Danilo S. Carvalho, and André Freitas. 2025. Interpreting token compositionality in LLMs: A robustness analysis. *arXiv preprint*. ArXiv:2410.12924 [cs] version: 2.

Jacob Andreas. 2019. Measuring Compositionality in Representation Learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Rim Assouel, Pietro Astolfi, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. 2024. OC-CLIP : Object-centric Binding in Contrastive Language-Image Pretraining. In *NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward*.

Rim Assouel, Pau Rodriguez, Perouz Taslakian, David Vazquez, and Yoshua Bengio. 2022. Object-centric Compositional Imagination for Visual Abstract Reasoning. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*.

Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. 2022. Logic Tensor Networks. *Artificial Intelligence*, 303:103649.

Renée Baillargeon and Julie DeVos. 1991. Object Permanence in Young Infants: Further Evidence. *Child Development*, 62(6):1227–1246.

Jasmijn Bastings, Marco Baroni, Jason Weston, Kyunghyun Cho, and Douwe Kiela. 2018. Jump to better conclusions: SCAN both left and right. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 47–55, Brussels, Belgium. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan

Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Elisabeth Camp. 2007. Thinking with Maps. *Philosophical Perspectives*, 21(1):145–182.

Patrick Cavanagh. 2021. The Language of Vision*. *Perception*, 50(3):195–215. Publisher: SAGE Publications Ltd STM.

François Chollet. 2019. On the Measure of Intelligence. *arXiv preprint*. ArXiv:1911.01547 [cs].

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*, 50 edition. The MIT Press.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical Foundations for a Compositional Distributional Model of Meaning. *arXiv preprint*. ArXiv:1003.4394 [cs].

Róbert Csordás, Kazuki Irie, and Juergen Schmidhuber. 2021. The Devil is in the Detail: Simple Tricks Improve Systematic Generalization of Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 619–634, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert Cummins. 1996. Systematicity. *The Journal of Philosophy*, 93(12):591–614.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022. The Paradox of the Compositionality of Natural Language: A Neural Machine Translation Case Study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175, Dublin, Ireland. Association for Computational Linguistics.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Charles Blundell, Philippe Beaudoin, Nicolas Heess, Michael Mozer, and Yoshua Bengio. 2021. Neural Production Systems. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual.*, pages 25673–25687.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is Winoground Hard? Investigating Failures in Visuolinguistic Compositionality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Leonidas A. A. Doumas, Guillermo Puebla, Andrea E. Martin, and John E. Hummel. 2022. A theory of relation learning and cross-domain generalization. *Psychological Review*, 129:999–1041. Place: US Publisher: American Psychological Association.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Joël Fagot and Carole Parron. 2010. Relational matching in baboons (Papio papio) with reduced grouping requirements. *Journal of Experimental Psychology: Animal Behavior Processes*, 36(2):184–193. Place: US Publisher: American Psychological Association.

Joël Fagot, Edward A. Wasserman, and Michael E. Young. 2001. Discriminating the relation between relations: The role of entropy in abstract conceptualization by baboons (Papio papio) and humans (Homo sapiens). *Journal of Experimental Psychology: Animal Behavior Processes*, 27(4):316–328. Place: US Publisher: American Psychological Association.

Jerome Feldman. 2013. The neural binding problem(s). *Cognitive Neurodynamics*, 7(1):1–11.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. 2025. Monitoring Latent World States in Language Models with Propositional Probes. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Jiahai Feng and Jacob Steinhardt. 2024. How do Language Models Bind Entities in Context? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jerry Fodor and Brian P. McLaughlin. 1990. Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, 35(2):183–204.

Jerry A. Fodor and Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1):3–71.

Gottlob Frege. 1892. *Über Sinn und Bedeutung*, 1. auflage edition. Zeitschrift für Philosophie und philosophische Kritik, Neue Folge. Pfeffer, Leipzig.

Gerald Gazdar, Evan Klein, Geoffrey K. Pullum, and Ivan A. Sag. 1985. *Generalized Phrase Structure Grammar*. Harvard University Press.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673. Publisher: Nature Publishing Group.

Adele Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, Chicago ; London. Series Title: Cognitive theory of language and culture.

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. 2020. On the Binding Problem in Artificial Neural Networks. *arXiv preprint*. ArXiv:2012.05208 [cs].

Robert F. Hadley. 1994. Systematicity in Connectionist Language Learning. *Mind & Language*, 9(3):247–272.

Per-Kristian Halvorsen and William A. Ladusaw. 1979. Montague's 'universal grammar': An introduction for the linguist. *Linguistics and Philosophy*, 3(2):185–223.

Jaakko Hintikka. 1979. Language-Games. In Esa Saarinen, editor, *Game-Theoretical Semantics: Essays on Semantics by Hintikka, Carlson, Peacocke, Rantala, and Saarinen*, pages 1–26. Springer Netherlands, Dordrecht.

Robert Huben, Hoagy Cunningham, Logan Riggs, Aidan Ewart, and Lee Sharkey. 2024. Sparse Autoencoders Find Highly Interpretable Features in Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wilhelm von Humboldt. 1836. *On Language: On the Diversity of Human Language Construction and Its Influence on the Mental Development of the Human Species*. Google-Books-ID: _UODbGlD4WUC.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality Decomposed: How do Neural Networks Generalise? *Journal of Artificial Intelligence Research*, 67:757–795.

Theo M. V. Janssen and Barbara H. Partee. 1997. Chapter 7 - Compositionality. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*, pages 417–473. North-Holland, Amsterdam.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, Honolulu, HI. IEEE.

Aravind K. Joshi. 2005. Tree-Adjoining Grammars. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, page 0. Oxford University Press.

jylin04, JackS, Adam Karvonen, and Can. 2024. OthelloGPT learned a bag of heuristics.

Ali Kazmi and Francis Jeffry Pelletier. 1998. Is Compositionality Formally Vacuous? *Linguistics and Philosophy*, 21(6):629–633. Publisher: Springer.

Najoung Kim and Tal Linzen. 2020. COGS: A Compositional Generalization Challenge Based on Semantic Interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Yeongbin Kim, Gautam Singh, Junyeong Park, Caglar Gulcehre, and Sungjin Ahn. 2023. Imagine the Unseen World: A Benchmark for Systematic Generalization in Visual World Models. *Advances in Neural Information Processing Systems*, 36:27880–27896.

Seijin Kobayashi, Simon Schug, Yassir Akram, Florian Redhardt, Johannes von Oswald, Razvan Pascanu, Guillaume Lajoie, and João Sacramento. 2024. When can transformers compositionally generalize in-context? *arXiv preprint*. ArXiv:2407.12275 [cs].

Brenden Lake and Marco Baroni. 2018. Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2873–2882. PMLR. ISSN: 2640-3498.

Kevin J. Lande. 2021. Mental structures. *Noûs*, 55(3):649–677.

Kevin J. Lande. 2024. Compositionality in perception: A framework. *WIREs Cognitive Science*, 15(6):e1691.

Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. 2024. Does CLIP Bind Concepts? Probing Compositionality in Large Image Models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1487–1500, St. Julian's, Malta. Association for Computational Linguistics.

Bingzhi Li, Lucia Donatelli, Alexander Koller, Tal Linzen, Yuekun Yao, and Najoung Kim. 2023a. SLOG: A Structural Generalization Benchmark for Semantic Parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3213–3232, Singapore. Association for Computational Linguistics.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yuxuan Li and James L. McClelland. 2022. Systematic Generalization and Emergent Structures in Transformers Trained on Structured Tasks. *arXiv preprint*. ArXiv:2210.00400 [cs].

Weiduo Liao, Ying Wei, Mingchen Jiang, Qingfu Zhang, and Hisao Ishibuchi. 2023. Does Continual Learning Meet Compositionality? New Benchmarks and An Evaluation Framework. *Advances in Neural Information Processing Systems*, 36:33499–33513.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. 2020. Object-Centric Learning with Slot Attention. In *Advances in Neural Information Processing Systems*, volume 33, pages 11525–11538. Curran Associates, Inc.

R Duncan Luce. 1996. The ongoing dialog between empirical science and measurement theory. *Journal of Mathematical Psychology*, 40(1):78–98.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. CREPE: Can Vision-Language Foundation Models Reason Compositionally? In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10910–10921, Vancouver, BC, Canada. IEEE.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

R. Thomas McCoy, Tal Linzen, Ewan Dunbar, and Paul Smolensky. 2020. Tensor Product Decomposition Networks: Uncovering Representations of Structure Learned by Neural Networks. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 277–278, New York, New York. Association for Computational Linguistics.

Kate McCurdy, Paul Soulos, Paul Smolensky, Roland Fernandez, and Jianfeng Gao. 2024. Toward Compositional Behavior in Neural Models: A Survey of Current Views. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9323–9339, Miami, Florida, USA. Association for Computational Linguistics.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. Language Models Implement Simple Word2Vec-style Vector Arithmetic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5030–5047, Mexico City, Mexico. Association for Computational Linguistics.

Arsenii Kirillovich Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. 2023. The ConceptARC Benchmark: Evaluating Understanding and Generalization in the ARC Domain. *Transactions on Machine Learning Research*.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent Linear Representations in World Models of Self-Supervised Sequence Models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30, Singapore. Association for Computational Linguistics.

nostalgebraist. 2020. interpreting GPT: the logit lens.

Victor V. Odouard and Melanie Mitchell. 2022. Evaluating Understanding on Conceptual Abstraction Benchmarks. *Proceedings of the AI Evaluation Beyond Metrics Workshop (IJCAI-2022)*.

Maya Okawa, Ekdeep Singh Lubana, Robert P. Dick, and Hidenori Tanaka. 2023. Compositional Abilities Emerge Multiplicatively: Exploring Diffusion Models on a Synthetic Task. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Chris Olah. 2023. Distributed Representations: Composition & Superposition.

Gustaw Opiełka, Hannes Rosenbusch, and Claire E. Stevenson. 2025. Analogical Reasoning Inside Large Language Models: Concept Vectors and the Limits of Abstraction. *arXiv preprint*. ArXiv:2503.03666 [cs].

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 39643–39666, Vienna, Austria. JMLR.org.

Barbara H. Partee. 2004. Compositionality. In *Compositionality in Formal Semantics*, pages 153–181. John Wiley & Sons, Ltd. Section: 7.

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251):20220041. Publisher: Royal Society.

Jean Piaget. 2013. *The Construction Of Reality In The Child*. Routledge, London.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Jake Quilty-Dunn, Nicolas Porot, and Eric Mandelbaum. 2023. The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, 46:e261.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Jacob Russin, Sam Whitman McGrath, Danielle J. Williams, and Lotem Elber-Dorozko. 2024. From Frege to chatGPT: Compositionality in language, cognition, and deep neural networks. *arXiv preprint*. ArXiv:2405.15164 [cs].

Lukas Schott, Julius von Kügelgen, Frederik Träuble, Peter Vincent Gehler, Chris Russell, Matthias Bethge, Bernhard Schölkopf, Francesco Locatello, and Wieland Brendel. 2022. Visual Representation Learning Does Not Generalize Strongly Within the Same Domain. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Prithviraj Sen, Breno W. S. R. de Carvalho, Ryan Riegel, and Alexander Gray. 2022. Neuro-Symbolic Inductive Logic Programming with Logical Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8212–8219. Number: 8.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. 2025. Open Problems in Mechanistic Interpretability. *arXiv preprint*. ArXiv:2501.16496 [cs].

Paul Soulos, Henry Conklin, Mattia Opper, Paul Smolensky, Jianfeng Gao, and Roland Fernandez. 2024. Compositional Generalization Across Distributional Shifts with Sparse Tree Operations. *Advances in Neural Information Processing Systems*, 37:112836–112863.

Mark Steedman. 2019. Combinatory Categorial Grammar. In András Kertész, Edith Moravcsik, and Csilla Rákosi, editors, *Current Approaches to Syntax*, A Comparative Handbook, pages 389–420. de Gruyter Mouton.

Stanley Smith Stevens. 1946. On the theory of scales of measurement. *Science*, 103(2684):677–680.

Kaiser Sun, Adina Williams, and Dieuwke Hupkes. 2023. The Validity of Evaluation Results: Assessing Concurrence Across Compositionality Benchmarks. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 274–293, Singapore. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? Probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Jonathan Thomm, Giacomo Camposampiero, Aleksandar Terzic, Michael Hersche, Bernhard Schölkopf, and Abbas Rahimi. 2024. Limits of Transformer Language Models on Learning to Compose Algorithms. *Advances in Neural Information Processing Systems*, 37:7631–7674.

Roger K. R. Thompson, David L. Oden, and Sarah T. Boysen. 1997. Language-naive chimpanzees (Pan troglodytes) judge relations between relations in a conceptual matching-to-sample task. *Journal of Experimental Psychology: Animal Behavior Processes*, 23(1):31–43. Place: US Publisher: American Psychological Association.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5228–5238, New Orleans, LA, USA. IEEE.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function Vectors in Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon M. Kleinberg, and Sendhil Mullainathan. 2024. Evaluating the World Model Implicit in a Generative Model. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Martin Wattenberg and Fernanda Viégas. 2024. Relational Composition in Neural Networks: A Survey and Call to Action. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Dag Westerståhl. 1998. On Mathematical Proofs of the Vacuity of Compositionality. *Linguistics and Philosophy*, 21(6):635–643. Publisher: Springer.

Zhengxuan Wu, Christopher D. Manning, and Christopher Potts. 2023. ReCOGS: How Incidental Details of a Logical Form Overshadow an Evaluation of Semantic Interpretation. *Transactions of the Association for Computational Linguistics*, 11:1719–1733.

Michael E. Young and Edward A. Wasserman. 1997. Entropy detection by pigeons: Response to mixed visual displays after same–different discrimination training. *Journal of Experimental Psychology: Animal Behavior Processes*, 23(2):157–170. Place: US Publisher: American Psychological Association.

Michael E. Young and Edward A. Wasserman. 2001. Entropy and variability discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1):278–293. Place: US Publisher: American Psychological Association.

Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. 2022. A Benchmark for Compositional Visual Reasoning. *Advances in Neural Information Processing Systems*, 35:29776–29788.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. RAVEN: A Dataset for Relational and Analogical Visual REasoNing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5317–5327.