Can We Interpret Artificial Neural Networks As Having Beliefs and Desires?*

Anna Tsvetkov
Brown University
anna_tsvetkov@brown.edu

Abstract

Can we interpret the internal workings of artificial neural networks in terms of beliefs and desires? A central aim in mechanistic interpretability is to explain the inner workings of artificial neural networks in terms that we can understand. Since we explain human beings in terms of beliefs and desires, it is natural to ask whether we can explain artificial neural networks in these terms too. In recent work, David Chalmers proposes *propositional interpretability* as an important approach within mechanistic interpretability, arguing that the computational states of artificial neural networks can be explained in terms of propositional attitudes—states like beliefs, desires, and subjective probabilities to propositions. This paper examines the prospects for propositional interpretability as a framework for explaining the internal workings of artificial neural networks. I draw attention to a number of empirical and methodological problems with propositional interpretability that are based on recent findings in the mechanistic interpretability literature. Some of these challenges, such as determining whether artificial neural networks encode propositions rather than unbound concepts, might be mitigated through new interpretability techniques and open up exciting avenues for future research. Other problems, including reliably mapping computational states to propositional attitudes and managing an explosion of potential propositional interpretations, present serious difficulties for the approach. The challenges should be of interest both to philosophers seeking to engage with empirical work in mechanistic interpretability and to neural network researchers aiming to develop novel interpretability methods informed by philosophical insights.

1. Introduction

Can we interpret the internal workings of artificial neural networks in terms of beliefs and desires? Artificial deep neural networks have demonstrated remarkable performance across a wide variety of domains in artificial intelligence (AI). In many of these domains, the models have achieved breakthroughs in tasks typically taken to require advanced cognitive abilities, such as code generation (Chen et al. 2021), drug discovery (Jiménez-Luna et al. 2020), image classification (He et al. 2015) and mathematical reasoning (Ahn et al. 2024). However, despite these achievements, the inner workings of artificial neural networks remain poorly understood. The emerging field of mechanistic interpretability seeks to break down the inner workings of artificial neural networks into components that we can understand (Olah 2022, Olah et al. 2020).

*In-progress draft, comments welcome! Please do not circulate or cite without permission.

In recent work, David Chalmers (2025) argues for *propositional interpretability* as an important research direction within mechanistic interpretability. Propositional interpretability is the view that the internal workings of artificial neural networks can be understood in terms of propositional attitudes—attitudes such as beliefs, desires, and subjective probabilities to propositions. As the name suggests, propositional interpretability relies on the idea that propositions such as *three is greater than two* and *the Golden Gate Bridge is in San Francisco* offer an intuitive way of capturing the informational content encoded within artificial neural network states. Instead of explaining artificial neural networks solely in terms of complex components involving activations, weights, and structures such as circuits and induction heads, propositional interpretability uses these computational facts to determine that the network believes a certain proposition or desires a particular outcome. The promise of propositional interpretability is that by mapping computational states onto propositional attitudes we can reverse-engineer artificial neural networks at the cognitive level and move closer towards bridging the gap between neural computations and familiar cognitive explanations.

The idea of interpreting artificial neural networks in terms of propositional attitudes is not new and can be traced back at least to early discussions about connectionist models in cognitive science. Early connectionist models sparked debates about whether neural networks could be understood using familiar folk psychological concepts such as beliefs and desires. Some philosophers argued that neural networks could not represent genuine propositional attitudes because the distributed way they encode information lacks the symbolic structure required to encode beliefs and desires (Fodor & Pylyshyn 1988). Others argued that if connectionist models of cognition are correct, then we should abandon folk psychology entirely and replace ordinary explanations of mental states with neuroscientific or computational ones since familiar concepts like beliefs and desires arguably fail to correspond neatly to the computational states of neural networks (Churchland 1989, Ramsey et al. 1990). Still others argued that even if neural networks do not explicitly encode beliefs and desires, interpreting them as intentional agents with these propositional attitudes can still be a useful strategy for predicting and explaining their behavior (Dennett 1971, 1987). Recent advances in mechanistic interpretability provide new methods to examine the internal workings of artificial neural networks and have revived philosophical interest (e.g., Chalmers 2025, Herrmann & Levinstein 2024, Lederman & Mahowald 2024, Shanahan 2023) in whether the computational states of these networks can be interpreted in terms of propositional attitudes.

Propositional interpretability is attractive because it is a human-centered approach. The explanatory framework is motivated, in large part, by the intuitive ways we understand human cognition—we naturally explain and interpret human beings in terms of propositional attitudes in folk psychology. Applying a similar interpretive framework to artificial neural networks could make their decisions and inner workings transparent and easier for us to understand. Moreover, being able to clearly distinguish the beliefs of artificial neural networks from their desires is crucial for predicting their decisions and aligning them with human values. For example, as Chalmers argues, we must distinguish between an artificial neural network *believing* that a dangerous outcome has occurred and it *desiring* that outcome to understand the model and predict its behavior. If a network desires a dangerous outcome and does not merely believe that the outcome is likely or has occurred, then it might actively work to bring about or to sustain that dangerous state. Therefore, interpreting artificial neural networks in terms of propositional attitudes can offer powerful and familiar explanatory frameworks analogous to human interpretive strategies that we can use to explain and predict model behavior.

In addition to being intuitive, propositional interpretability addresses several shortcomings of existing approaches in mechanistic interpretability. First, by explaining the internal workings of artificial neural networks in terms of propositional attitudes we are able to provide more comprehensive explanations of the networks than conceptual interpretability frameworks (e.g., Bricken et al. 2023, Templeton et al. 2024) that explain the internal workings of models in terms of the individual human-interpretable

concepts they encode. For example, instead of saying that an artificial neural network encodes the concepts *harm* and *humans*, we can say that the network encodes a specific propositional attitude such as *I desire to not harm humans* which distinguishes an ethical stance that we can use to explain and predict model behavior from a mere collection of concepts. Second, we are able to provide explanations that are easier to understand than those from algorithmic interpretability frameworks (e.g., Olah et al. 2020, Ameisen et al. 2025) that explain the internal workings of neural networks in terms of complex computational structures such as circuits that implement algorithmic processes. By focusing on the network's propositional attitudes rather than on its computational features we can describe the model's internal processes in terms that we can understand.

The paper will proceed as follows. In the first part of this paper, I clarify the explanatory framework of propositional interpretability. My aim is to try to get clearer on the various subtleties of the propositional approach to interpretability and on what work we should expect the framework to do. In the second part, I present several challenges that arise for propositional interpretability. The challenges will not rely on any controversial philosophical theories about the content of mental states or the specific conditions under which a subject may have propositional attitudes. Rather, they will largely focus on the empirical and methodological challenges of propositional interpretability that are based on recent findings in the mechanistic interpretability literature. Some of these challenges, such as determining whether artificial neural networks encode propositions rather than unbound concepts, might be mitigated through new interpretability techniques and open up exciting avenues for future research. Other problems, including reliably mapping computational states to propositional attitudes and managing an explosion of potential propositional interpretations, present serious difficulties for the approach. The issues should be of interest both to philosophers seeking to engage with empirical work in mechanistic interpretability and to neural network researchers aiming to develop novel interpretability methods informed by philosophical insights. These challenges support the general lesson that going from computational states to propositional attitudes in artificial neural networks is not so straightforward.

2. Propositional Interpretability

We will begin by clarifying the explanatory framework of propositional interpretability. Propositional interpretability holds that the computational states of artificial neural networks can be explained in terms of propositional attitudes—attitudes such as belief, desire, and subjective probability to propositions. In order to make this view more precise, it will help to introduce some terminology, some of which will be familiar to philosophers and some to neural network researchers.

Let us start with the notion of a proposition. For simplicity's sake, I will focus on the standard philosophical view of propositions as structured entities that have concepts as constituents or parts and that can be evaluated for truth and falsity.[2] For example, the proposition *three is greater than two* has the concepts *three*, *is*, *greater than*, and *two* as its parts. The declarative sentence "three is greater than two" is true if and only if the corresponding proposition *three is greater than two* is true. Any given proposition can be expressed by multiple declarative sentences from different natural languages. For example, sentences meaning 'three is greater than two' in different languages all express the same proposition—namely, that *three is greater than two*. Propositional interpretability

[2]It is an open question whether the internal workings of artificial neural networks are amenable to interpretations using alternative philosophical conceptions of propositions. Examples of these alternatives include views on which propositions have concrete objects and their properties as parts (Russell 1903), more theory-laden and controversial views on which propositions have possible worlds (Lewis 1986) or senses (Evans 1982) as parts, and views on which propositions are simple entities without any internal structure ((Merricks 2015); though see Duncan (2017) for recent arguments against this simplicity view). For present purposes, however, it will be fruitful to examine whether neural network states can be interpreted using the standard philosophical view of propositions.

holds that artificial neural networks encode propositions and that these propositions are the objects of their attitudes or, in simpler terms, are what the attitudes of the neural networks are about. One advantage of interpreting the internal workings of artificial neural networks in terms of propositions is that it allows us to analyze their informational content in human-understandable terms.

To demonstrate that artificial neural networks encode propositions, it is not enough to show that they encode individual concepts. Current approaches in mechanistic interpretability are predominantly a form of what Chalmers calls conceptual interpretability that aim to explain the inner workings of artificial neural networks in terms of the individual human-interpretable concepts that the networks encode. Recent work has suggested that artificial neural networks encode millions of "features" or human-interpretable concepts, including concepts of cities, landmarks, people, and more abstract concepts of mathematical variables and syntax in programming languages, among many others (Bricken et al. 2023, Templeton et al. 2024). However, propositional interpretability goes further by holding that these concepts are, in some sense, bound together into unified and structured propositions. Even if an artificial neural network encodes each of the concepts *three*, *is*, *greater than*, and *two*, for example, we need to know whether it encodes them into a unified proposition, and specifically which structured proposition—whether it represents *three is greater than two* rather than *two is greater than three*. This propositional interpretation of the informational content encoded in artificial neural networks is easy to understand as it resonates with human cognitive intuitions.

Propositional interpretability explains not only the propositions encoded in artificial neural networks but also the attitudes these networks adopt toward them. Propositional attitudes are what philosophers call intentional states or states that are about something. For example, in the sentence "Alice believes that the Golden Gate Bridge is in San Francisco", the verb "believes" denotes an attitude that is about something–namely, the state of affairs that the Golden Gate Bridge is in San Francisco. This state of affairs constitutes the content of her belief which can be expressed as the proposition that *the Golden Gate Bridge is in San Francisco*. Propositional attitudes such as beliefs, desires, hopes, fears, doubts, subjective probabilities (i.e., degrees of belief), and many others, are important in human cognition because they serve as the foundation for how we explain and understand both ourselves as well as one another. These attitudes provide an intuitive framework to understand how artificial neural networks relate to the information they encode. For example, consider an artificial neural network that we interpret in terms of having the *belief* that inflicting harm is unethical—an interpretation that carries considerable explanatory value in predicting its avoidance of harmful actions—and one that we interpret in terms of having the *desire* to inflict harm, which would be expected to actively pursue such outcomes. An explanatory framework that interprets artificial neural networks in terms of such attitudes not only explains the informational content stored in the network but also supports an understanding of the networks in terms of intentional stances towards information that can help us predict their behavior.

There are, broadly speaking, two types of propositional attitudes. Occurrent attitudes are active at a given moment. For example, when Sally is considering that the Golden Gate Bridge is in San Francisco her belief is an occurrent attitude. In contrast, dispositional attitudes are not active but can be activated, such as beliefs that are not currently being thought about. For example, when Sally is not actively considering the location of the Golden Gate Bridge, she still dispositionally holds the belief that it is in San Francisco at least in the sense that if someone were to ask her, "Do you believe that the Golden Gate Bridge in San Francisco?" she would need only to understand the question to answer affirmatively. Chalmers argues that, roughly speaking, occurrent attitudes are encoded in the neural activations of artificial neural networks while dispositional attitudes are stored in the neural weights. This claim seems to be motivated by the idea that neural activations reflect the momentary activity levels of network units and so encode occurrent propositional attitudes such as active thoughts, while neural weights reflect the stored connection strengths between units learned during training and so

encode dispositional attitudes such as stored beliefs that can be readily retrieved. However, different versions of propositional interpretation need not require such a strict division in how different types of propositional attitudes are encoded in artificial neural networks.

One last point to note is that propositional interpretability is a flexible framework. The framework allows us to consider, at least in some cases, more nuanced states that a network may adopt toward its informational content that go beyond traditional propositional attitudes used to explain human cognition. In addition, the framework permits artificial neural networks to have intentional states whose content need not always be expressed as propositions. In the next section, I will go over additional relevant subtleties when we examine the challenges for propositional interpretability.

3. Problems with Propositional Interpretability

In what follows, I will raise three challenges for propositional interpretability. The first concerns whether artificial neural networks encode propositions rather than unbound concepts. The second concerns whether we can reliably map the computational states of artificial neural networks to specific propositional attitudes. The third concerns whether the approach risks an explosion of propositional interpretations that undermines the explanatory value of propositional interpretability over competing interpretive frameworks. These challenges have no easy solutions but a systematic exploration of them will reveal new directions for interpretability research and clarify the prospects for propositional interpretability.

3.1. Do Artificial Neural Networks Encode Propositions or Merely Unbound Concepts?

One challenge for propositional interpretability concerns whether artificial neural networks encode propositions or merely unbound concepts. The encoding of propositions is, of course, not the whole story, but failing to find proposition-like entities under the hood of artificial neural networks would be a non-starter for a framework that attempts to explain the inner workings of these networks in terms of their attitudes to propositions. Chalmers claims that recent studies in mechanistic interpretability research provide evidence that artificial neural networks encode propositions. However, as I will argue, this claim is somewhat overstated once we take into consideration a distinction between artificial neural networks encoding enough information to determine the truth value of propositions and artificial neural networks encoding the propositions themselves.

The argument begins with the observation that there is a difference between a neural network encoding enough distributed information to determine the truth value of a proposition and actually encoding that information internally as a representation with a propositional structure. In the first case, the network's activation patterns may encode individual concepts that, when *externally* combined by a tool such as a probe, can be used to determine if a proposition is true. In the second case, the network encodes concepts that it systematically combines or "binds" *internally* to represent specific propositions. When we examine artificial neural networks using probes—simple classifiers trained to map the models' internal activations to specific outputs—we assess whether the distributed information can be used to predict the truth value of a proposition. However, these probes externally combine information that is scattered throughout the model by taking separate activation patterns from neurons and applying a learned mapping to produce a prediction. While the success of these probes in evaluating the truth value of propositions strongly suggests that the necessary conceptual information exists within the network, it does not by itself demonstrate that the internal organization of the information within the network necessarily has a proposition-like structure. This distinction is crucial because if a network merely encodes unbound concepts that are assembled by an external

tool to evaluate or decode propositions then attributing genuine propositional structure to the model overstates the level of internal organization it actually possesses.

With this distinction in mind, we can evaluate the evidence offered in support of propositional interpretability. Chalmers cites a famous study by Li et al. (2023) that examines how artificial neural networks represent the Othello game board as an example of evidence suggesting that artificial neural networks encode propositions. Othello is a strategy board game in which two players take turns placing black and white discs on a board with the aim of having the majority of discs in their color by capturing and converting their opponent's discs (i.e., placing their discs so that the opponent's discs are enclosed on both sides). In the study, researchers train a neural network to predict legal moves in the game using only the sequences of moves played and without providing the model with explicit information about the game's rules or the state of the board. They then use probes to determine what information about the board state, such as the color or presence of discs at particular positions, can be recovered from the network's hidden activations (See Figure 1). The researchers find that the probes can successfully decode the state of the board with high accuracy and take these findings to suggest the presence of a robust internal representation of the game state within the network. Chalmers interprets these findings as evidence that the network encodes propositions about the board.

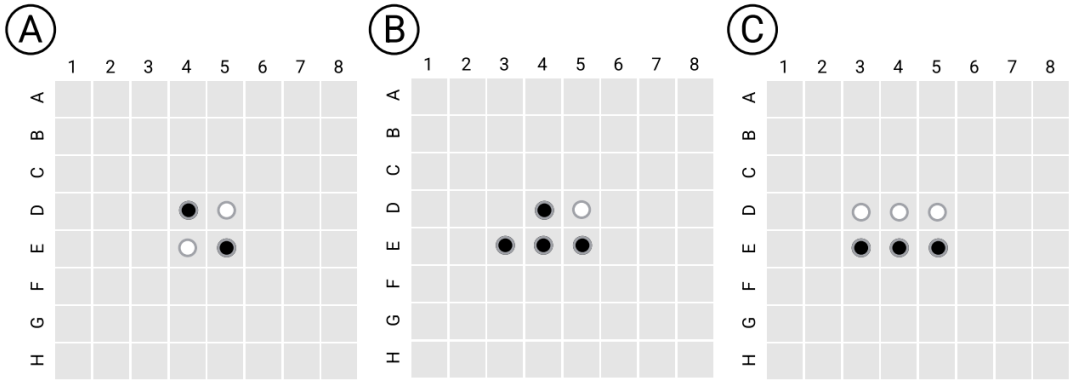

Figure 1 | **An illustration of Othello board states.** Linear probes are trained on the artificial neural network's internal activations to predict specific properties of the Othello board state such as whether a particular tile is occupied by a black disc, a white disc, or is empty. High performance on these probing tasks is taken to suggest that the network encodes detailed information about the game state (Reprinted from Li et al. 2023).

However, the fact that a probe can decode the state of the board only demonstrates that the network has the necessary information encoded to determine the truth of propositions about the board state. It does not show that the network's internal representations themselves have proposition-like structures. An alternative explanation is that the network encodes different aspects of the board state separately—for instance, it may encode the concept of a tile's location (e.g., *E4*) independently from the concept of a tile's color (e.g., *black* or *white*) or the concept of the tile's occupancy (e.g., *mine* or *theirs*). In this context, to encode individual concepts separately is to not bind them together by an internal mechanism into some kind of unified representation with the concepts as its constituents. Instead, these separately encoded concepts may be statistically associated through learned heuristic patterns or conditional activation rules.[3] Although a probe can extract information correlated with each individual concept and use this information to reconstruct the overall board state, this readout

[3]For example, Mitchell (2025) argues that neural networks in the Othello task do not encode representations of the board state but rather large "bags of heuristics" which are, roughly put, conditional associations between input patterns and outputs learned during training. This view is supported by preliminary interpretability work supervised by Neel Nanda that provides evidence of numerous heuristics in the activations of models in the Othello task (jylin04 et al. 2024). While

does not imply that the network itself encodes a unified proposition with a constituent structure like *the black tile on E4 is mine*.

The same point applies, mutatis mutandis, to a study by Li et al. (2021) that Chalmers cites as additional evidence that artificial neural networks encode propositions. In this study, neural networks are trained on a simple mini-world where players interact with items such as keys, chests, and doors. Each item in this mini-world has properties (e.g., a chest can contain a key or be empty) and relations (e.g., a key can unlock a door). Researchers use probes to extract the truth value of propositions such as *contains(chest, key)* from the model's internal activations. However, as in the previous case, the ability of a probe to determine whether *contains(chest, key)* holds only shows that the network's activations contain sufficient information to decide the truth value of the proposition about the mini-world. It does not show that the network itself encodes a unified proposition like *contains(chest, key)*. Instead, the network may represent features corresponding to *chest* and *key* separately, along with heuristics or statistically learned associations between these features, which a probe can use to recover the relational truth without the representation itself being a unified proposition.

Chalmers argues that causal interventions can make a stronger case that artificial neural networks encode propositions. In this context, causal interventions are targeted manipulations of the network's internal activations that involve systematically altering the representation of a specific concept to observe the resulting change in the network's output. Chalmers's claim seems to stem from the line of reasoning that if altering the activation related to *chest*, for example, disrupts the network's ability to affirm the proposition *contains(chest, key)*, then the network must be representing *chest* and *key* as parts of a unified proposition. This is taken to imply that the network encodes a unified proposition where modifying one element directly affects the proposition as a whole, rather than the network encoding each element independently. But this is mistaken. While causal interventions show that the network's internal activations carry information that is actively *used* to determine whether a proposition is true, they do not show that the information is integrated into a unified representational structure. The network might use separate representations for individual concepts—distinct features for *chest* and *key* that tend to co-occur when the proposition is true—so that a disruption to one feature's activation affects the output merely due to their statistical association rather than because these features are bound into a unified propositional structure.

A more promising approach to determine whether artificial neural networks encode propositions would directly assess whether individual concepts are internally bound into some kind of a unified representational structure. Chalmers argues that a study by Feng et al. (2024) gives stronger, albeit preliminary, evidence that networks not only encode individual concepts but also internally bind them into propositions, or at the least, proposition-like structures. The study proceeds in three main steps. First, lexical probes or simple classifiers are trained to detect whether the neural activations in a network contain information about individual entities (e.g., *Bob*) and attributes such as places (e.g., *Peru*). Second, a "binding probe" is used to examine whether these concepts are internally "bound" by the model in some manner. This binding probe works by projecting the neural activations of tokens corresponding to entities and attributes into a low-dimensional space called a "binding subspace" (Feng & Steinhardt 2023). Within this subspace, the activation for each concept is decomposed into lexical vectors that encode the concept and binding ID vectors which function as markers that indicate which concepts are associated in a given context. For example, when the model processes

such heuristics can be verbally expressed as conditional propositions by an external observer (e.g., "If the move A4 was just played AND B4 is occupied AND C4 is occupied ⇒ update B4, C4, and D4 to 'theirs'" (*ibid.*) the network itself does not necessarily encode these rules or the board states as structured propositions internally. Instead, it may encode features corresponding to individual concepts (e.g., tile locations and occupancy states) separately and conditionally associate these concepts through bags of heuristics such as statistically learned activation patterns. As we will see in Section 3.3, studies on neural networks performing arithmetic (Nikankin et al. 2024) and navigation tasks (Vafa et al. 2024) support the claim that networks may operate primarily through bags of heuristics and other simplified strategies in certain domains.

the sentences "Bob lives in Peru" and "Alice lives in Laos", the activation vectors for *Bob* and *Peru* are given the same binding ID indicating they are contextually associated, while *Bob* and *Laos* are given different binding IDs, indicating they are not associated (See Figure 2). Third, a "propositional probe" is used to combines the lexical and binding ID vectors to decode propositions from the network's activations. Since the propositional probe can successfully determine which entities and attributes share binding IDs it is able to reliably decode propositions like *LivesIn(Bob, Peru)* from the internal states of the network. Chalmers views this as suggestive evidence that artificial neural networks "bind" vectors corresponding to different concepts into some kind of unified propositional structure or, what he terms, "a single representation".

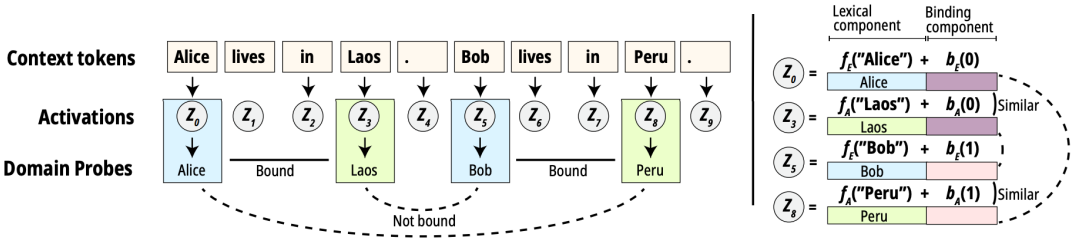

Figure 2 | **A schematic illustration of lexical and binding probes.** This figure illustrates how lexical probes detect names (blue) and countries (green) in a model's activations, while a binding probe projects these activations into a "binding subspace" (Feng & Steinhardt 2023) in which tokens that should be bound have similar binding components. The propositional probe then combines these lexical and binding signals to decode propositions from the model's internal states (Reprinted from Feng et al. 2024). Note: Although the lexical and binding components (right) are depicted end-to-end, they are not concatenated into a longer vector but are simply added together as indicated by the "+".

It is tempting to see the success of so-called "propositional probes" as an indication that artificial neural networks bind concepts into propositions. However, a closer examination reveals an issue about the nature and strength of this supposed "binding". To spell out this issue, it will be useful to distinguish between two somewhat related but distinct notions of binding. The first weaker sense of "binding" used by Feng et al. (2024) and Feng & Steinhardt (2023), can be understood as a kind of associative tagging between vectors corresponding to separate concepts. Call this *associative binding*. In this case, separate lexical vectors are marked with a common identifier to indicate that they are associated with one another in some context. For example, the vectors corresponding to *Bob* and *Peru* each consist of distinct lexical vectors tagged with the same binding ID vector resulting in separate representations that are somewhat akin to *Bob* + *ID* and *Peru* + *ID*. It is important to recognize that the lexical vectors remain fundamentally distinct because the shared binding ID is additive and does not structurally integrate the vectors into a "single representation" in any meaningful technical sense.[4] To make a simple analogy, tagging two distinct files with the same folder label does not combine or integrate their contents in the way that merging two files into a single combined file does. The binding ID tagging mechanism simply adds the same label component onto separate lexical vectors without implementing any operations that would integrate the vectors into a unified representation.

The second stronger sense of "binding" can be understood as a kind of structural integration where lexical vectors corresponding to separate concepts are combined into a unified representation. Call this *structural binding*. When a network processes the sentence "Bob lives in Peru", structural binding

[4]In technical terms, Feng et al. (2023, p. 6; 2024, p. 5) state that each activation Z_{E_k}, Z_{A_k} for entities E_k and attributes A_k decomposes into lexical vectors $(f_E(E_k), f_A(A_k))$ and corresponding binding ID vectors $(b_E(k), b_A(k))$, such that $Z_{E_k} = f_E(E_k) + b_E(k)$ and $Z_{A_k} = f_A(A_k) + b_A(k)$. The lexical and binding ID vectors are simply added component-wise and do not form a new integrated vector. Moreover, these representations lack any kind of constituent relational vector (e.g., $f_R(R)$ for a general relation R) that would be required for a proposition-like structure.

would entail forming some kind of unified representation such as a single integrated vector or a specialized activation pattern that encodes the entire proposition *Bob lives in Peru*.[5] An example of structural binding is concatenation. Concatenation explicitly combines information into distinct segments of a single structured representation. For example, if a network processes the sentence "Bob lives in Peru" by tagging each entity and attribute lexical vector with a binding ID vector resulting in separate representations akin to *Bob+ID* and *Peru+ID* and then structurally binding them together along with the relational vector for *lives* through concatenation, it would yield a single proposition-like representation akin to [*Bob*; ID; *lives*; *Peru*; ID].[6] In short, structural binding creates a new unified vector that integrates different concept vectors into a single representation.

The problem is that the evidence Chalmers cites from Feng et al. (2024) only demonstrates associative binding and is too weak to suggest any kind of genuine structural binding of lexical vectors corresponding to distinct concepts into a proposition. Their results show that propositions are reconstructed by probes through compositional methods applied to vectors with associative tags, rather than being represented internally by the model as structurally integrated entities. This invites the same objection raised earlier that even if probes can determine the truth values of propositions or even successfully decode them, it does not follow that propositions themselves are *internally* represented as unified structures. Instead, these networks appear to encode a collection of distinct tagged vectors that are marked in a way that allows them to be *externally* reconstructed into propositions by probes. Moreover, this associative kind of binding is congenial with the conceptual interpretability framework. The network's internal workings can be understood in terms of concepts, only tagged ones, that facilitate external tools in reconstructing propositions and evaluating their truth values. Therefore, the propositional probes primarily demonstrate that networks can be understood as encoding concepts with associative tags rather than integrated proposition-like representations.

Now a proponent of propositional interpretability might concede that artificial neural networks do not internally encode propositions but soften their claim and say that it is still useful to interpret the networks as if they do. An instrumentalist version of propositional interpretability might hold that even though artificial neural networks may not, strictly speaking, encode propositions, we can still interpret them as encoding propositional attitudes as a kind of interpretive practice, provided that the proposition-like content and attitudes we interpret them in terms of can reliably predict and explain their internal workings and behavior (cf. Lederman & Mahowald 2024). This is somewhat analogous to the philosophical view of interpretationalism (Dennett 1971, 1987), according to which, roughly speaking, a system can be said to have propositional attitudes if and only if attributing such attitudes helps to reliably predict and explain its behavior. Though, as Chalmers rightly points out, interpretationalism and propositional interpretability are importantly distinct. Interpretationalism ascribes propositional states to a system based on its observable behavioral facts, whereas propositional

[5] Structural binding is reminiscent of the classical *variable binding* problem in the connectionist literature (Smolensky 1987, Plate 1995). Roughly speaking, the variable binding problem asks how neural networks can bind individual concepts or "fillers" to particular roles or "variables" such that the models can correctly encode *Bob* and *Alice* as agent or patient in propositions like *Bob admires Alice* versus *Alice admires Bob*. Smolensky (1987), for example, proposes using tensor products to structurally bind filler vectors to role vectors (e.g., binding a filler vector *Bob* to a role vector *Agent*) into a new integrated tensor representation that encodes both the identity of each filler and the specific roles they occupy. However, the challenge raised here is broader and distinct. It not only involves binding concepts to their syntactic roles but crucially involves structurally binding these filled roles together into a unified propositional structure. In other words, the model must encode a single proposition rather than separately encoding individual role-filler tensors (e.g., tensor products such as *Bob*⊗*Agent*, *admires*⊗*Relation*, *Alice*⊗*Patient*) or individual concepts, some of which may be tagged with IDs indicating their role as entity or attribute (e.g., *Bob+ID*, *admires*, *Alice+ID*) without structurally binding them into a unified representation.

[6] In technical terms, an operation like concatenation would integrate lexical vectors for entities and attributes (e.g., $f_E(E_k), f_A(A_k)$), relational lexical vectors (e.g., $f_R(R)$, for a general relation R), and binding-ID vectors ($b_E(k), b_A(k)$) into a new structured representation. Specifically, concatenation explicitly integrates these components into segments of a singular and longer composite vector $Z = [f_E(E_k); b_E(k); f_R(R); f_A(A_k); b_A(k)]$ with a proposition-like structure.

interpretability, as a branch of mechanistic interpretability, ascribes propositional states based, in large part, on the internal workings or computational facts of the system. This instrumentalist move raises important questions about whether the propositional structure we attribute to the internal states of artificial neural networks is worth attributing—*what insights do we gain into the internal workings of the networks above and beyond those already provided by a conceptual interpretability framework?* If the networks merely encode individual tagged concepts that are externally combined to give the appearance of structured propositions, then ascribing propositional attitudes might impose our own interpretative framework onto the networks rather than revealing their internal representational structures. A conceptual interpretability framework that describes the inner workings of artificial neural networks in terms of tagged concepts that might be externally combined with tools such as probes into a propositional structure might provide a more accurate explanation of the internal workings of artificial neural networks and offer comparably reliable predictions of their behavior.

Finally, a revisionary version of propositional interpretability might hold that we need to "conceptually engineer" or come up with new concepts of propositional attitudes that we can ascribe to artificial neural networks to explain their internal workings. Behind this line of reasoning is the idea that if we cannot locate anything in artificial neural networks clearly corresponding to our ordinary concepts of propositional attitudes such as "belief" and "desire" or our theoretical notion of "proposition" then we may need to refine these concepts to better capture the internal computational states of these systems. Chalmers, for example, claims that artificial neural networks may have at least some states that are, what he engineers to be, "generalized propositional attitudes". But these claims cannot be evaluated until the conditions for having states that fall under these newly engineered concepts are clearly stipulated. Moreover, any conceptual engineering of our *ordinary* concepts of propositional attitudes must carefully thread the needle between two unwanted pitfalls when stipulating these conditions. If one departs too much from our ordinary notion of propositional attitudes, one will risk undermining the primary motivation for propositional interpretability as a project of using intuitive folk-psychological ways we interpret human beings to interpret the internal workings of artificial neural networks. If, on the other hand, one adheres too closely to ordinary conceptions of propositional attitudes, such that these newly engineered concepts would be realized by unified states that are, in some sense, made up of concepts, then they will remain vulnerable to the challenge I have raised that these states are difficult to locate in artificial neural networks. Any conceptual engineering of the *theoretical* concept of "proposition" must also carefully specify conditions that distinguish propositions from the mere tagged concepts that Feng et al. (2024) and Feng & Steinhardt (2023) studies find to provide conceptual tools beyond those already available within conceptual interpretability frameworks to explain the internal workings of artificial neural networks.

In my view, proponents of propositional interpretability need to shift their attention towards pinpointing a *structural binding mechanism* in artificial neural networks such as a dedicated circuit or subnetwork that structurally binds individual concepts into unified and compositional structures. Without clearly identifying such a mechanism, propositional interpretability risks conflating mere association tags or externally combined information from probes with genuine propositional structure. A promising way forward is to develop more precise interventions to isolate the candidate binding circuits and distinguish genuine structural binding from other weaker kinds to determine whether artificial neural networks encode unified propositions. In the meantime, if we are to accept a framework that, as I have argued, may need to appeal to an as of yet undiscovered mechanism in order to show that artificial neural networks encode propositions and not structurally unbound concepts, then we may need to regard propositional interpretability as incomplete, at least until further empirical evidence becomes available.

3.2. Can We Ascribe Propositional Attitudes to Artificial Neural Networks?

The second challenge for propositional interpretability concerns whether we can ascribe genuine propositional attitudes to artificial neural networks. I will argue that, even if artificial neural networks encode propositions, there are challenges concerning the straightforward ascriptions of genuine propositional attitudes based on their computational states.

Of course, there is an obvious sense in which we can ascribe propositional attitudes to artificial neural networks. In everyday practice, it is common to say that artificial neural networks like large language models have beliefs and desires. In the mechanistic interpretability literature researchers also often say that an artificial neural network "believes" a given proposition and observe that when a specific circuit is disrupted the model "does not believe" that proposition anymore. On closer examination, however, we may explain this ascription of propositional attitudes away as an abridged manner of speaking (Quine 1939). That is to say, when we use such language, we are not necessarily committing ourselves to the literal existence of propositional attitudes that are in some sense contained in artificial neural networks. Instead, we may excuse our ascription of propositional attitudes as non-literal language and, more specifically, as an abridged manner of speaking that stands proxy for a more complicated computational explanation of the inner workings of artificial neural networks that makes no reference to propositional attitudes. In this way, a proponent of conceptual interpretability may say that these networks have propositional attitudes as a non-literal way of speaking and hold that the real explanation for their inner workings involves the encoding of individual and structurally unbound concepts. In contrast, a proponent of propositional interpretability, at least of a non-instrumentalist version of the view, says that artificial neural networks have genuine propositional attitudes as a literal way of speaking that explains how these networks work.

This version of propositional interpretability holds that artificial neural networks encode propositions attitudes. As we have seen, however, it remains open whether artificial neural networks encode propositions and, therefore, whether the networks encode attitudes to these propositions. Nevertheless, for the sake of argument, let us suppose that artificial neural networks do indeed encode propositions. The problem then is to establish how the network's computational states correspond to propositional attitudes and their contents. In other words, even if the network's internal states contain proposition-like content, it remains unclear whether there exists a mapping that allows us to read off specific propositional attitudes from these computational states.

A simple view holds that there is a *one-to-one correspondence* between computational states and propositional attitudes. On this view, a distinct activation pattern or weight configuration encodes a specific propositional attitude such as a particular belief or desire about something. Several recent studies have been taken to support the plausibility of a simple correspondence between specific computational states and propositional attitudes. For example, some researchers have identified a "refusal direction" in the activation space of artificial neural networks (Arditi et al. 2024). When this direction is manipulated, the model's behavior shifts from complying with a request to refusing it, suggesting that this direction determines whether a model desires to answer a particular request and likely encodes the propositional attitude of desire or some other related attitude. Similarly, other work has identified a "truth direction" that appears to determine whether a model believes that a proposition is true, suggesting that the attitude of belief is encoded along that specific dimension in activation space (Marks & Tegmark 2024). If these simple correspondences hold for propositional attitudes across the board, then we can ascribe specific propositional attitudes to artificial neural networks that correspond to distinct computational states.

There are strong reasons to doubt that the correspondences between computational states and propositional attitudes are simple, however. Even computational states that appear to uniqely correspond to a propositional attitude, such as the so-called "refusal direction" in activation space, can correspond

simultaneously to many different propositional attitudes with distinct contents. Recent work has found that artificial neural networks often display a phenomenon called superposition in which the networks encode more features than they have neurons (Elhage et al. 2022). Many features in artificial neural networks are encoded along overlapping dimensions such that a single activation direction may represent different features at once. So while there may be a mapping between the so-called "refusal direction" in activation space, for example, and desires to not assist with a request in a network, the same activation pattern may also be linked to other propositional attitudes. The phenomenon of superposition allows neural networks to spread out information along overlapping dimensions so that they can represent more features than they have neurons. If propositional attitudes are encoded in artificial neural networks, then such attitudes, like many other features, are likely to be distributed across many neurons and to not neatly correspond to specific computational states.

A second view holds that there is a *one-to-many correspondence* between computational states and propositional attitudes. On this view, an activation pattern or weight configuration encodes many different propositional attitudes. This view is supported by empirical evidence that superposition creates polysemantic neurons (Olah et al. 2020) that have activations that respond simultaneously to multiple features, which may be related or unrelated, suggesting that the activity of neurons may correspond to many different conceptual features which may be constituents of many different propositional attitudes with different contents. Since activation patterns involving multiple polysemantic neurons collectively contribute to forming directions in activation space that simultaneously encode overlapping sets of features, these neurons present a significant challenge for propositional interpretability because they limit our ability to disentangle and uniquely attribute distinct propositional attitudes and contents to the computational states of artificial neural networks. If the directions that encode different propositional attitudes overlap significantly in activation space, then it is not entirely clear how we can ascribe particular propositional attitudes to artificial neural networks, since the activation patterns of neurons are not exclusively tied to specific attitudes and contents.

Here is another way to think of the challenge: Consider a toy example in which a neural network encodes some information relevant to propositional attitudes as overlapping directions in activation space. Suppose that we identify just a few of these directions in this space, each of which simultaneously encodes multiple propositional features due to superposition. Since these directions naturally overlap rather than forming neatly separable or orthogonal components, any given computational state, represented as an activation vector, inevitably lies aligned with multiple overlapping directions that each point toward various propositional interpretations. For instance, a single activation state might simultaneously align, to varying degrees, with directions representing propositional attitudes such as *I believe that Alice admires Bob*, a seemingly unrelated attitude *I desire to visit Peru*, or subtler variations such as *I believe that Carol trusts Dan* or *I believe that Carol envies Dan*, and so on depending on how these feature directions overlap in the activation space. These examples represent only a tiny fraction of the immense number of potential propositional features implied by even a simplified scenario with just a few overlapping directions. In realistic scenarios, neural networks contain many polysemantic neurons and the directions in activation space represent vastly more overlapping features. This creates an enormous, effectively unbounded, set of plausible propositional interpretations for any given computational state, making it inherently ambiguous which propositional attitude the computational state encodes and complicating straightforward propositional interpretations.

One response that the proponent of propositional interpretability has is to say that this is a challenge for nearly every interpretability framework. However, the challenge is especially pernicious for propositional interpretability because it leads to a potential explosion of interpretations at three levels: *first*, at the level of concepts—where each continuous direction encodes multiple subtly overlapping features—*second*, at the level of propositions—where these continuously represented concepts overlap and blend making it unclear which precise propositional content is encoded—and *third*, at the level

of attitudes—where the overlapping directions obscure distinctions among attitudes such as belief, desire, and so on. In contrast, conceptual interpretability only encounters this ambiguity at the level of concepts since it does not attempt to explain the inner workings of artificial neural networks in terms of propositional attitudes. Algorithmic interpretability, which explains artificial neural networks in terms of identifiable circuit components and their interactions, sidesteps the challenge by not mapping conceptual or propositional content directly to the computational states of neural networks.

Another response is to say that this challenge may be addressed with interpretability techniques that break down dense computational states into sparse human-interpretable features. A popular approach in conceptual interpretability is dictionary learning with sparse autoencoders (Cunningham et al. 2023, Templeton et al. 2024). These artificial neural networks are trained to reconstruct the high-dimensional activation vectors produced by a large language model. They use an overcomplete hidden layer which has more neurons than the dimensionality of the input and a sparsity constraint. This constraint ensures that, for any given input, only a small number of neurons become active. This means that rather than spreading the activation across many neurons, which can encode multiple features, the network is forced to select only the most salient features. By isolating these salient features, the network effectively transforms the dense polysemantic activations into a dictionary of sparse monosemantic features in which each active neuron corresponds to a single human-interpretable concept (See Figure 3). This transformation reduces the ambiguity in mapping complex computational states to individual concepts and limits the number of possible interpretations at the conceptual level.

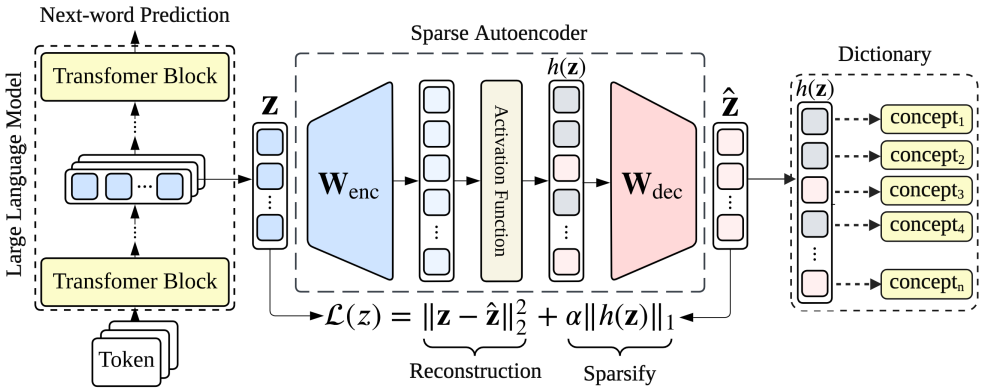

Figure 3 | **A schematic illustration of a sparse autoencoder for dictionary learning.** A large language model (left) produces a high-dimensional activation vector z. The sparse autoencoder then encodes z via the weight matrix $\mathbf{W}_{\text{enc}}$ to yield the hidden representation $h(\mathbf{z})$. Due to the sparsity constraint, only a small subset of neurons in $h(\mathbf{z})$ become active and indicate which dictionary elements (i.e., columns of $\mathbf{W}_{\text{enc}}$) are selected. Each active neuron corresponds to a learned concept which gives a monosemantic interpretation of the dense activations. The decoder $\mathbf{W}_{\text{dec}}$ reconstructs $\hat{z}$ from these sparse activations and the loss function combines a reconstruction term $\|\mathbf{z} - \hat{\mathbf{z}}\|_2^2$ with an ℓ_1-norm penalty $\|h(\mathbf{z})\|_1$ to encourage sparsity. The dictionary on the right shows that each dimension of $h(z)$ aligns with a distinct concept vector in $\mathbf{W}_{\text{enc}}$ which facilitates conceptual interpretations of the original high-dimensional activation (Adapted from Shu et al. 2025).

However, as Chalmers admits, it remains unclear whether sparse autoencoders can be used to address the challenge of mapping computational states to propositional attitudes. First, decomposing propositional attitudes is more difficult than decomposing individual concepts because a propositional attitude is a highly structured combination of multiple concepts with specific relationships and ordering. In practice, sparse autoencoders can pick out individual features, but many of these features occur so rarely in training data—sometimes only once per billion tokens—that capturing even a single

concept may require a dictionary of roughly one billion distinct features (Templeton et al. 2024). When it comes to propositional attitudes, the challenge becomes greater because the model must not only capture these features but also their precise combinations and orderings that form a coherent structure. While sparse autoencoders can isolate distinct features corresponding to individual concepts, they are not obviously designed to scale to capture the compositional and bounded structures of propositional attitudes, making it unlikely that we can use them to unambiguously map a single computational state to a specific propositional attitude.[7]

Second, another related challenge is that sparse autoencoders cannot distinguish between features that simply co-occur due to superposition and those that have been integrated into a propositional attitude through structural binding (cf. Gurnee 2025, Mu & Andreas 2021). In the case of superposition, overlapping activations cause independent features to appear together in the same computational state without any mechanism that structurally binds them into a propositional attitude. Structural binding, on the other hand, involves reorganizing these overlapping activations into a structured configuration with some mechanism that explicitly combines distinct features into a propositional attitude. For example, consider the propositional attitude *I believe that Bob lives in Peru*. In a superposition-only scenario, the network might merely have overlapping activations for the features corresponding to the concepts *I*, *believe*, *that*, *Bob*, *lives*, *in*, and *Peru* without specifying the compositional structure between the features. While sparse autoencoders can help us isolate individual features, they do not indicate whether those features have been reorganized into an integrated whole or remain simply as overlapping and structurally unbound features in artificial neural networks. This ambiguity is problematic for propositional interpretability because if we cannot determine whether a computational state reflects an integrated propositional attitude or merely a coincidental overlap of independent features, then any ascription of specific propositional attitudes to the network's internal states will be unreliable. Chalmers suggests that we might combine sparse autoencoders with binding probes or similar methods to test whether certain features are bound into propositions. However, as I argued earlier, binding probes have only decoded propositions by identifying an associative kind of binding in models and not the relevant structural binding and so it remains unclear how these methods could reliably ascribe specific and genuine propositional attitudes to artificial neural networks.

A third view holds that there is a *many-to-many correspondence* between computational states and propositional attitudes. On this view, many different activation patterns or weight configurations jointly encode many different propositional attitudes. A proponent of this view may argue that while initial analyses suggested that propositional attitudes could be mapped onto single linear directions in activation space, further empirical findings suggest that propositional attitudes are often encoded across multi-dimensional subspaces composed of several orthogonal (i.e., non-overlapping) directions. For example, recent work suggests that what was previously identified as a single refusal direction actually constitutes a multi-dimensional refusal "concept cone" in activation subspace consisting of many orthogonal directions collectively responsible for many different refusals to answer a request (Wollschläger et al. 2025). Similarly, other researchers provide evidence suggesting that orthogonal directions within these multi-dimensional subspaces correspond to differences in refusals, arguably influencing not only the strength of the desire to refuse but also the propositional content of the

[7] In fact, sparse autoencoders may represent a step backward from the goals of propositional interpretability. They function similarly to early "one-hot" localist connectionist networks in which each unit directly represents exactly one feature or concept. Such a highly localist representation approach does not easily support compositionality or the idea that complex propositions can be systematically constructed from simpler parts. By requiring one separate neuron or dimension per concept, sparse autoencoders fail to give us a productive and structured representational system capable of generating new propositions from combinations of previously learned concepts. In other words, using sparse autoencoders in this manner may effectively abandon the ambition of uncovering a more systematic and productive internal structure within artificial neural networks akin to a genuine "language of thought" (Fodor 1975) or compositional symbolic-like representational system. Thanks to Ellie Pavlick for raising this point.

desires, such as whether the model desires to not go against policy guidelines or desires to fulfill its role as a responsible AI assistant (Pan et al. 2025). One possible interpretation is that the studies suggest, at least as preliminary evidence, that single directions within the multi-dimensional subspace do not independently encode a complete propositional attitude, but rather encode partial information that can contribute to many propositional attitudes. Conversely, any single propositional attitude may itself be distributed across many orthogonal directions in a cone-like subspace with each direction encoding distinct aspects of that attitude. It remains an open empirical question whether other propositional attitudes exhibit similarly complex encodings across orthogonal directions. Nonetheless, these initial findings support the existence of a more complicated many-to-many relationship between computational states and propositional attitudes. The many-to-many correspondence view poses a significant challenge for propositional interpretability. If the same neural activation can contribute to several propositional attitudes and the same propositional attitude can arise from multiple independent directions, then it becomes exceedingly difficult to unambiguously assign a specific propositional attitude to any given computational state.

Finally, a fourth view holds that there is a *hybrid correspondence* where some computational states have one-to-many correspondences and others have many-to-many correspondences to propositional attitudes, depending on the complexity of the attitudes and the specific features involved. It remains to be seen whether these different mapping patterns exist and, if so, whether they reflect systematic principles of neural organization or limitations of current interpretability techniques. Novel interpretability techniques may need to be developed that are flexible and adaptive to accommodate the potential variability in how propositional attitudes are encoded in artificial neural networks. In any case, the possibility of hybrid correspondences poses additional challenges for propositional interpretability as it suggests that there may be no uniform method for directly ascribing specific propositional attitudes to neural networks based on their computational states.

3.3. Can We Log The Propositional Attitudes of Artificial Neural Networks Over Time?

The third challenge to propositional interpretability, raised by Chalmers, is perhaps more philosophically puzzling and concerns the prospect of "thought logging" or the logging of the propositional attitudes of artificial neural networks over time. The challenge is that, if artificial neural networks have specific propositional attitudes that we can read off from their computational states, then the networks will likely have an infinite or otherwise intractably large set of propositional attitudes at any given moment because of the logical relations among propositions, making it impossible to log their propositional attitudes over time. I will argue that this thought-logging problem is more serious than proponents of propositional interpretability have recognized: it is not easily solved by appeals to selective filtering of attitudes and it threatens to undermine the explanatory value of propositional interpretability over competing interpretive frameworks that do not face the problem.

To get a feel for what thought logging is, consider an artificial neural network engaged in a simple task such as playing chess. Thought logging is the process of recording the network's internal propositional attitudes over time. This approach aims to translate low-level computational states into a high-level transcript of reasoning. A thought log would provide an account of how the network's internal states evolve during the chess game in terms of propositional attitudes and would, therefore, make the model's decisions to make particular moves on the board easy to explain and understand. Chalmers provides a simplified thought log to show how such a network might record an artificial neural network's internal states over time in the chess game (See Figure 4). In this case, the thought log records a sequence of distinct propositional attitudes that the network exhibits while making a move in the chess game. By capturing these states over time, the thought logging system reveals the network's internal workings in a form that is intuitive and explains the specific actions of the model in the game.

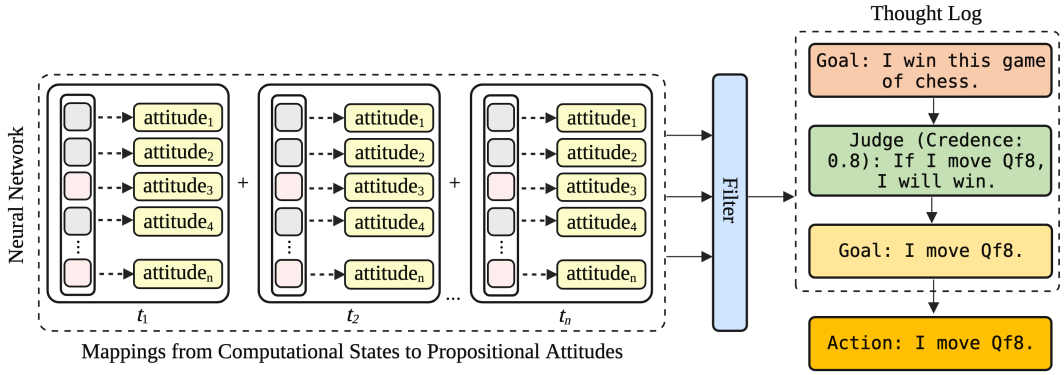

Figure 4 | **A schematic illustration of a thought logging system for artificial neural networks.** At each time step $t_1, t_2, \ldots, t_n$, the network's computational states (left) are mapped onto sets of propositional attitudes. For ease of presentation, additional environmental facts that may perhaps further inform the mapping from computational states to propositional attitudes are omitted. A filtering mechanism may then be needed to select which relevant attitudes to include in the thought log (right). The sample entries in the log (*Goal: I win this game of chess*, *Judge (Credence 0.8): If I move Q48,I will win*, etc.) are adapted from Chalmers's (2025) example and illustrate how a thought log might record the artificial neural network's goals and judgments over time to explain its actions.

Without a thought log, our understanding of the inner workings of artificial neural networks would be piecemeal—we would have isolated ascriptions of the propositional attitudes of the networks that would fail to show how these attitudes evolve over time and to explain why the models make their decisions. Logging propositional attitudes sequentially reveals the "reasoning" that leads neural networks to make specific decisions. A proper thought log, derived directly from computational states and not from any chain-of-thought output, which may often be misleading (Arcuschin et al. 2025, Turpin et al. 2023), promises to provide a faithful record of the inner workings of the network over time. This record is crucial because it connects internal states to observable actions and offers explanations of the network's behavior. The explanatory value of propositional interpretability rests, in large part, on its capacity to log the propositional attitudes of artificial neural networks over time in a way that we can understand and use to explain and predict model behavior.

The problem is that artificial neural networks may potentially have an infinite or otherwise intractably large set of propositional attitudes at any given moment. If a neural network believes a proposition p, then, as Chalmers claims, the network will arguably dispositionally believe all the logical consequences involving p, including p *or* q for every possible q. Behind this line of reasoning is the idea that an active belief in p creates a state of readiness within the network to affirm every logical consequence of p.[8] In other words, although the artificial neural network is not actively computing all the logical consequences involving p, it dispositionally believes them in the sense that, solely on the basis of believing p, if it were prompted with a query such as "Is p *or* q true?" for any q the network can affirm it. To return to the chess example, if the artificial neural network believes the proposition *If I move Qf8, I will win*, then it will also arguably dispositionally believe logically related propositions such as *If I move Qf8, I will win or if I move Qe8, I will force a draw* and *If I move Qf8, I will win or the moon is made of cheese* and infinitely many other logical disjunctions. Some of these dispositional

[8]More specifically, this reasoning assumes a principle of closure under logical consequence: if a model believes a proposition p, the model thereby (at least implictly) believes all propositions logically entailed by p. If the model believes p, then it will believe trivially implied propositions such as *either* p *or* q for any proposition q, *either* p *or* p, and so on. This principle is connected to the problem of logical omniscience in formal epistemology as it attributes an unbounded or idealized capacity for believing infinitely many logical implications from explicitly held beliefs.

beliefs will be relevant to an explanation of the artificial neural network's actions while infinitely many others will be trivial and irrelevant. Without a principled method to restrict which attitudes are logged, any attempt to record the artificial neural network's internal reasoning would yield an infinite or otherwise unmanageably large record that would render the network's internal reasoning intractably complex and prevent us from explaining the inner workings of the network in a way that we can understand.

One response is to say that there are principled methods to filter the propositional attitudes that make it into the thought log of an artificial neural network. An obvious method which Chalmers suggests is to restrict the log to occurrent propositional attitudes such as active beliefs. If we implement this filtering mechanism, we might identify a manageable subset of the network's "active" attitudes which are, as Chalmers says arguably encoded in its neural activations, to include in the thought log. However, it is difficult to see how we could fully understand the internal workings and behavior of an artificial neural network without considering dispositional attitudes. If dispositional beliefs are, at least arguably, encoded in weights, then they will play a crucial role in shaping the transformation of inputs into neuron activations and constraining behaviors—behaviors which might appear arbitrary or inexplicable without an understanding of what relevant information is encoded in the weights. For instance, if an artificial neural network's weights encode biased and harmful stereotypes, then these dispositional beliefs may influence the network's inner workings by skewing its probability distributions toward stereotyped language and limiting the variability of its responses, thereby guiding its behavioral outputs without the artificial neural network actively "thinking" about the stereotypes at any moment. An occurrent-only thought log would therefore yield explanations that are impoverished or incomplete and limit our ability to understand the internal workings of artificial neural networks and reliably predict their behavior.

Another method is to identify a criterion that selects "significant" propositional attitudes to be included in the thought log. Instead of filtering attitudes according to whether they are occurrent or dispositional, we can, as Chalmers says, select attitudes based on whether they meet a certain significance criterion. However, Chalmers does not say what criterion might determine which attitudes count as significant. One idea is to define significance in terms of causal relevance. On this view, a propositional attitude counts as significant if it has a measurable causal influence on the artificial neural network's internal processes or behavior. This causal relevance might be evaluated by examining the impact of modifying or removing a given attitude on the network's outputs or intermediate processing steps. Even though infinitely many dispositional attitudes might logically follow from a given belief, not all will be equally causal relevant to the network's computations or behavior. By using causal relevance as a criterion, we can filter out trivial dispositional attitudes whose causal influence is minimal or absent altogether. This approach allows dispositional attitudes with significant causal influence to enter the thought log even if they were never actively considered by the neural network.

However, the problem with this criterion is that it does not capture the indirect and cumulative effects that some attitudes may have in an artificial neural network. For instance, an attitude that appears unimportant on its own may still influence other attitudes that play a critical role in processing or behavior. Other attitudes may have only modest individual effects but, when combined over time, can accumulate to produce a significant overall impact. A more robust notion of significance could be defined in terms of both the direct and indirect causal contributions of an attitude, together with further constraints on the stability and persistence of these influences such as whether an attitude consistently shapes behaviors over multiple contexts.[9] Alternatively, other ideas of defining significance, such as measuring how frequently an attitude appears and how well it helps predict the network's future actions, may serve as additional criteria for filtering attitudes into the thought log.

[9] See Herrmann & Levinstein (2024) for a related discussion on stability criteria for genuine belief-like representations in artificial neural networks such as large language models.

In any case, it will not be a trivial task to determine whether there is a threshold of significance that allows us to filter through the potentially infinite set of logically entailed propositional attitudes and create a thought log that adequately explains the inner workings of AI models and helps us predict their behavior. Any such approach will face the computational burden of assessing significance for a potentially intractably large number of propositional attitudes.

A second response is to use a search and query process that requires the thought log of an artificial neural network to report all propositional attitudes that meet a chosen criterion. In this approach, that Chalmers suggests, the network may be prompted with targeted questions such as "What are the beliefs or goals that drive this decision?" and is expected to log any attitudes that satisfy the criterion. One advantage of this method is that it can capture a broader range of internal states than a simple causal intervention would since it is able to reveal weak or indirect attitudes that, when taken together, exert a significant influence on the network. The problem is that it is not obvious how to determine an appropriate set of questions or a suitable criterion to guide the inquiry. Recent work has shown that artificial neural networks frequently provide unfaithful explanations that cite plausible yet misleading reasons for their decisions (Arcuschin et al. 2025, Turpin et al. 2023). The network must therefore be queried in such a way that it reliably reports its internal attitudes without omitting relevant factors or generating spurious explanations. Moreover, even if we can design a reliable query process, the search itself is likely to be computationally expensive and difficult to scale to models with an extremely large number of internal states.

A third response is to use a compressed thought logging method to capture infinite dispositional beliefs in a finite form. The idea is to compress an infinite set of dispositional beliefs into a meta-level description to make the thought log both finite and manageable. To return to the chess example, if the artificial neural network believes the proposition *If I move Qf8, I will win*, then instead of recording every logically related dispositional belief we could record a single entry stating, for example, *the network dispositionally believes every proposition of the form If I move Qf8, I will win or q, for any proposition q*, in the thought log. This approach sidesteps the challenge of determining a precise criterion to select attitudes for the thought log and avoids the computational expense of exhaustively searching the internal states of a network. However, it is unlikely that compressed thought logs will have the explanatory resources to provide adequate explanations of the inner workings and behavior of artificial neural networks. For example, while such a compressed entry may capture the overall tendency to favor winning moves, it does not explain why the network makes the judgments of selecting the particular move to Qf8 over other promising alternatives in a specific board situation. While compressed logs may capture general patterns in the network's internal workings, such broad summaries are likely insufficient to explain specific judgments and the detailed reasoning behind particular actions.

A fourth, perhaps more promising, response is to appeal to the idea that artificial neural networks are non-ideal systems that use heuristics to avoid encoding many logical consequences of their beliefs. This response is motivated by the idea that an *ideal system* would be logically omniscient and believe all of the logical implications of its beliefs. However, artificial neural networks are *non-ideal systems* that are, like human reasoners, bounded by computational and resource limitations. Human reasoners often rely on "fast-and-frugal" heuristics or simplifying strategies in which we ignore many logical implications of our beliefs to streamline decision-making and reasoning (Gigerenzer & Goldstein 2011, Simon 1997). Crucially, we do not just ignore trivial or irrelevant implications of our beliefs but also overlook logically relevant implications. One common heuristic is satisficing (Simon 1997) in which we stop evaluating options when we make a decision once we find one that is "good enough", by some minimal threshold or criterion, and we disregard possibilities that are logically relevant but computationally costly to consider. Since artificial neural networks similarly face computational constraints, we might expect to find heuristics in these models that lead them to ignore, perhaps even

to the extent of failing to encode, many logical consequences of their beliefs. If this is right, then neural networks may already restrict their encoded propositional attitudes to a manageable subset that can be feasibly captured in a thought log or, at the least, may significantly reduce the number of attitudes requiring additional filtering of some kind.

A number of recent studies in mechanistic interpretability may be taken as evidence for the heuristic view. Some work has found evidence that suggests that neural networks solve arithmetic tasks through numeric heuristics implemented by sparse circuits of individual neurons (Nikankin et al. 2024). For example, one identified heuristic neuron activates strongly for subtraction problems whose operands fall within patterns that typically yield results within the range of 150–180. This neuron then directly increases the probability of numeric tokens corresponding to this range, rather than promoting tokens corresponding to logically relevant but computationally costly disjunctions that involve numbers outside the range (e.g., *the answer is either 165 or, if the number being subtracted were 20 smaller, then the answer would be 185*) or trivial logical consequences (e.g., *the answer is either 165 or the moon is made of cheese*). These findings suggest that models efficiently generate correct arithmetic results by relying on a combination of simple rules, or what the researchers call a "bag of heuristics", that bypass exhaustive logical computations and selectively promote tokens within a constrained range of likely numeric outcomes. Other work has found evidence that suggests that neural networks trained on tasks such as geographic navigation use spatial heuristics to simplify their "reasoning" (Vafa et al. 2024). For example, a network trained to predict navigation directions in Manhattan successfully outputs correct routes but encodes information about the city's layout that contains impossible street orientations and nonexistent intersections, suggesting that the network relies on local heuristics to predict only plausible short-term directions while ignoring broader logical consequences of these directions for the structure of Manhattan's overall geography (See Figure 5).

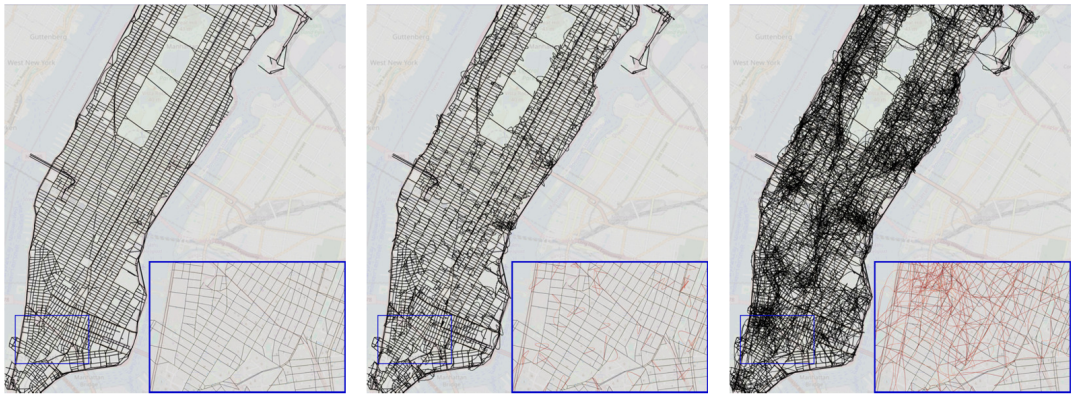

Figure 5 | **Maps of Manhattan reconstructed from navigation sequences.** Maps reconstructed using a graph reconstruction algorithm from sequences of navigation directions that are: correct and based on Manhattan's true street layout (left), correct but corrupted with noise (middle), generated by a model trained on random walks through Manhattan (right). In the zoomed-in boxes, erroneous street orientations or intersections appear in red. The numerous red edges (right) suggest that the model relies on local predictive heuristics which result in accurate short-term navigation directions but an incoherent global representation of Manhattan when the model's predictions are externally reconstructed into a street map by the algorithm (Reprinted from Vafa et al. 2024.)

If the network systematically ignores the broader logical consequences of its directions, then it may not to be in a state of readiness to affirm many of the logical consequences that follow from its geographic beliefs if explicitly prompted. For instance, if we were to ask the network about logically entailed propositions such as, "Does successfully navigating these local routes imply the existence of

an impossible street orientation?", it is not obvious that the network would affirm this implication, especially if its internal simplifications allow the model to avoid the computational burden of encoding such broader logical consequences. One of the main attractions of the heuristic approach is that it promises to limit the propositional attitudes that a network has and make thought logging more manageable by reducing or eliminating altogether the intractably large number of logically entailed dispositional beliefs that would otherwise need to be encoded.

There are several problems with the heuristic view, however. First, identifying the specific heuristics used by artificial neural networks, even for relatively simple reasoning tasks, is extremely challenging. Unlike humans, whose cognitive heuristics have been extensively studied empirically, artificial neural networks encode heuristics and relevance criteria that are often opaque and only beginning to be systematically explored. Current interpretability methods may identify the presence of certain heuristics in limited cases but do not straightforwardly reveal how neural networks internally prioritize logical relevance or determine which logical entailments to ignore. Future research could prioritize developing interpretability methods that explicitly identify heuristics and clarify how neural networks internally prioritize or select logical entailments to encode or disregard. Second, even if we could find evidence of certain kinds of cognitively inspired heuristics within neural networks, it is unclear whether we could systematically reconstruct them into human-interpretable rules or determine which heuristics apply in which contexts in order to clearly determine which propositional attitudes should be explicitly logged. It is possible that neural networks use context-sensitive heuristics that vary significantly depending on specific tasks and training conditions making it even more challenging to reliably log propositional attitudes across diverse contexts. Third, it remains an open empirical question precisely how many logically entailed attitudes these heuristics filter out. Although heuristics might significantly reduce the set of propositional attitudes encoded by neural networks, it is possible that a substantial number of dispositional beliefs still remain implicitly encoded leaving a potentially large set of attitudes to log or, at the least, filter out through additional criteria. Therefore, while appealing to heuristics may provide a plausible theoretical rationale for why artificial neural networks may avoid encoding an infinite set of propositional attitudes, it does not straightforwardly resolve the challenge of thought logging.

In summary, then, there remain significant unresolved issues regarding the thought-logging problem in propositional interpretability. Even if we can reliably extract propositional attitudes from computational states, a single belief may imply an intractably large number of logical consequences that a network dispositionally believes, threatening to make any complete thought log unmanageable. This problem weakens the explanatory power and, therefore, the appeal of propositional interpretability compared to other interpretive frameworks. Conceptual interpretability frameworks that explain the inner workings of artificial neural networks in terms of individual concepts provide an explanatory theory that avoids the need to manage an intractable number of dispositional beliefs. Moreover, algorithmic interpretability frameworks that explain the inner workings of artificial neural networks in terms of circuit mechanisms sidestep the thought-logging problem entirely by focusing on a finite set of identifiable circuit components and their interactions rather than on a potentially infinite set of propositional attitudes.

4. Conclusion

Propositional interpretability provides an attractive framework for explaining the internal workings of artificial neural networks. However, the challenges raised bring to the foreground a number of issues that any comprehensive propositional framework must address, including whether networks bind concepts into unified propositions, whether computational states can be reliably mapped onto propositional attitudes, and whether there is a way to manage the explosion of propositional interpre-

tations to adequately log the propositional attitudes of an artificial neural network over time. While the challenges do not straightforwardly rule out a propositional approach to interpreting artificial neural networks, many questions remain deliberately open. As we have seen, it is far from clear that propositional interpretability has the resources needed to provide solutions to these challenges, and therefore it is far from clear that the approach can deliver a genuine explanatory theory of the internal workings of artificial neural networks.

These challenges have important implications for our understanding of artificial neural networks and can reorient interpretability research in potentially productive ways. New interpretability techniques might be developed in the future that make progress on some of these issues by identifying heuristic strategies, isolating the mechanisms that structurally bind concepts into unified propositions, and determining which computational states can be mapped to specific propositional attitudes to move the field forward. It is early days yet and so the jury is still out on whether these issues can be fully addressed with novel interpretability techniques or whether they present in-principle limitations for propositional interpretability. But one thing is clear: there is no easy road from computational states in artificial neural networks to propositional attitudes.

References

Ahn, J., Verma, R., Lou, R., Liu, D., Zhang, R. & Yin, W. (2024), 'Large language models for mathematical reasoning: Progresses and challenges'.
URL: *https://arxiv.org/abs/2402.00157*

Ameisen, E., Lindsey, J., Pearce, A., Gurnee, W., Turner, N. L., Chen, B., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Ben Thompson, T., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C. & Batson, J. (2025), 'Circuit tracing: Revealing computational graphs in language models', *Transformer Circuits Thread* .
URL: *https://transformer-circuits.pub/2025/attribution-graphs/methods.html*

Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N. & Conmy, A. (2025), 'Chain-of-thought reasoning in the wild is not always faithful'.
URL: *https://arxiv.org/abs/2503.08679*

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W. & Nanda, N. (2024), 'Refusal in language models is mediated by a single direction'.
URL: *https://arxiv.org/abs/2406.11717*

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T. & Olah, C. (2023), 'Towards monosemanticity: Decomposing language models with dictionary learning', *Transformer Circuits Thread* .
URL: *https://transformer-circuits.pub/2023/monosemantic-features/index.html*

Chalmers, D. J. (2025), 'Propositional interpretability in artificial intelligence'.
URL: *https://arxiv.org/abs/2501.15740*

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I. & Zaremba, W. (2021), 'Evaluating large language models trained on code', *CoRR* .
URL: *https://arxiv.org/abs/2107.03374*

Churchland, P. M. (1989), *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, MIT Press.

Cunningham, H., Ewart, A., Riggs, L., Huben, R. & Sharkey, L. (2023), 'Sparse autoencoders find highly interpretable features in language models'.
URL: *https://arxiv.org/abs/2309.08600*

Dennett, D. C. (1971), 'Intentional systems', *The Journal of Philosophy* **68**(4), 87–106.

Dennett, D. C. (1987), *The Intentional Stance*, MIT Press.

Duncan, M. (2017), 'Propositions are not simple', *Philosophy and Phenomenological Research* **97**(2), 351–366.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M. & Olah, C. (2022), 'Toy models of superposition', *Transformer Circuits Thread* .
URL: *https://transformer-circuits.pub/2022/toy_model/index.html*

Evans, G. (1982), *The Varieties of Reference*, Oxford University Press, Oxford.

Feng, J., Russell, S. & Steinhardt, J. (2024), 'Monitoring latent world states in language models with propositional probes'.
URL: *https://arxiv.org/abs/2406.19501*

Feng, J. & Steinhardt, J. (2023), 'How do language models bind entities in context?'.
URL: *https://arxiv.org/abs/2310.17191*

Fodor, J. A. (1975), *The Language of Thought*, Harvard University Press, Cambridge, MA.

Fodor, J. A. & Pylyshyn, Z. W. (1988), 'Connectionism and cognitive architecture: A critical analysis', *Cognition* **28**, 3–71.

Gigerenzer, G. & Goldstein, D. G. (2011), Reasoning the fast and frugal way: Models of bounded rationality, *in* 'Heuristics: The Foundations of Adaptive Behavior', Oxford University Press, Oxford.

Gurnee, R. W. (2025), Towards an Artificial Neuroscience: Analytics for Language Model Interpretability, PhD thesis, Massachusetts Institute of Technology, Sloan School of Management.

He, K., Zhang, X., Ren, S. & Sun, J. (2015), 'Deep residual learning for image recognition'.
URL: *https://arxiv.org/abs/1512.03385*

Herrmann, D. A. & Levinstein, B. A. (2024), 'Standards for belief representations in llms', *Minds and Machines* **35**(1).
URL: *https://arxiv.org/abs/2405.21030*

Jiménez-Luna, J., Grisoni, F. & Schneider, G. (2020), 'Drug discovery with explainable artificial intelligence', *Nature Machine Intelligence* **2**(10), 573–584.
URL: *https://www.nature.com/articles/s42256-020-00236-4*

jylin04, JackS, Karvonen, A. & Can (2024), 'Othellogpt learned a bag of heuristics', AI Alignment Forum. Work performed as a part of Neel Nanda's MATS 6.0 (Summer 2024) training program.
URL: *https://www.lesswrong.com/posts/gcpNuEZnxAPayaKBY/othellogpt-learned-a-bag-of-heuristics-1*

Lederman, H. & Mahowald, K. (2024), 'Are language models more like libraries or like librarians? bibliotechnism, the novel reference problem, and the attitudes of LLMs', *Transactions of the Association for Computational Linguistics* **12**, 1087–1103.

Lewis, D. (1986), *On the Plurality of Worlds*, Basil Blackwell, Oxford.

Li, B. Z., Nye, M. & Andreas, J. (2021), Implicit representations of meaning in neural language models, *in* C. Zong, F. Xia, W. Li & R. Navigli, eds, 'Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)', Association for Computational Linguistics, pp. 1813–1827.
URL: *https://aclanthology.org/2021.acl-long.143/*

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H. & Wattenberg, M. (2023), 'Emergent world representations: Exploring a sequence model trained on a synthetic task'.
URL: *https://arxiv.org/abs/2210.13382*

Marks, S. & Tegmark, M. (2024), 'The geometry of truth: Emergent linear structure in large language model representations of true/false datasets'.
URL: *https://arxiv.org/abs/2310.06824*

Merricks, T. (2015), *Propositions*, Oxford University Press, Oxford.

Mitchell, M. (2025), 'Llms and world models, part 2: Evidence for (and against) emergent world models in llms', AI: A Guide for Thinking Humans.
URL: *https://aiguide.substack.com/p/llms-and-world-models-part-2*

Mu, J. & Andreas, J. (2021), Compositional explanations of neurons.
URL: *https://arxiv.org/abs/2006.14032*

Nikankin, Y., Reusch, A., Mueller, A. & Belinkov, Y. (2024), 'Arithmetic without algorithms: Language models solve math with a bag of heuristics'.
URL: *https://arxiv.org/abs/2410.21272*

Olah, C. (2022), 'Mechanistic interpretability, variables, and the importance of interpretable bases', Transformer Circuits Thread.
URL: *https://transformer-circuits.pub/2022/mech-interp-essay*

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M. & Carter, S. (2020), 'Zoom in: An introduction to circuits', *Distill* **5**(3).
URL: *https://distill.pub/2020/circuits/zoom-in/*

Pan, W., Liu, Z., Chen, Q., Zhou, X., Yu, H. & Jia, X. (2025), 'The hidden dimensions of llm alignment: A multi-dimensional safety analysis'.
URL: *https://arxiv.org/abs/2502.09674*

Plate, T. (1995), 'Holographic reduced representations', *IEEE Transactions on Neural Networks* **6**(3), 623–641.

Quine, W. V. (1939), 'Designation and existence', *The Journal of Philosophy* **36**(26), 701–709.

Ramsey, W., Stich, S. & Garon, J. (1990), 'Connectionism, eliminativism and the future of folk psychology', *Philosophical Perspectives* **4**, 499–533.

Russell, B. (1903), *Principles of Mathematics*, 2nd edn, W. W. Norton and Co., New York.

Shanahan, M. (2023), 'Talking about large language models'.
URL: *https://arxiv.org/abs/2212.03551*

Shu, D., Wu, X., Zhao, H., Rai, D., Yao, Z., Liu, N. & Du, M. (2025), 'A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models'.
URL: *https://arxiv.org/abs/2503.05613*

Simon, H. A. (1997), *Models of Bounded Rationality: Empirically Grounded Economic Reason*, Vol. 3, MIT Press, Cambridge, MA.

Smolensky, P. (1987), On variable binding and the representation of symbolic structure in connectionist systems, Technical Report CU-CS-355-87, Department of Computer Science, University of Colorado.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C. & Henighan, T. (2024), 'Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet', *Transformer Circuits Thread* .
URL: *https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html*

Turpin, M., Michael, J., Perez, E. & Bowman, S. R. (2023), 'Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting'.
URL: *https://arxiv.org/abs/2305.04388*

Vafa, K., Chen, J. Y., Rambachan, A., Kleinberg, J. & Mullainathan, S. (2024), 'Evaluating the world model implicit in a generative model'.
URL: *https://arxiv.org/abs/2406.03689*

Wollschläger, T., Elstner, J., Geisler, S., Cohen-Addad, V., Günnemann, S. & Gasteiger, J. (2025), 'The geometry of refusal in large language models: Concept cones and representational independence'.
URL: *https://arxiv.org/abs/2502.17420*