

*****Preprint of paper accepted at *Philosophical Issues******

Corresponding Author: cameron.buckner@ufl.edu

Second Thoughts about Chains-of-Thought: Self-talk, Transparency, and Artificial Reason

Cameron Buckner and Arjun Krishna Kumar

Abstract:

The frontier of AI is being pushed forward by “Large Reasoning Models” (LRMs) that self-generate long textual “Chains-of-Thought” (CoTs) before answering user queries. There has been intense interest in the potential that these CoTs—which appear to reveal processes of reasoning reported in natural language sentences—may render transparent the computations that these models use to generate their answers. A large body of research has thus been directed at the question of whether CoTs are “faithful” to the model’s underlying computations, a question which has been answered with mixed results. Here, we point out that this research tradition has largely been behaviorist in character—with reasoning processes identified by a particular input-output profile. By contrast, philosophy and cognitive science have largely considered reasoning to be an intensional phenomenon, with reasoning processes bearing a many-to-one relationship to input-output profiles, defined in part by the way that agents construe their evidence as supporting their conclusion. Here, we articulate three ways in which CoTs could be intensionally faithful—but are pessimistic on the prospects of satisfying those notions of faithfulness with current methods (including by supplementing with mechanistic interpretability methods). We end by concluding that, so long as we do not overintellectualize the sense in which agents are sensitive to their own construals of their evidence, intensional notions of faithfulness may be useful to near-term AI research by fostering the development of new forms of flexible artificial epistemic agency and metacognition.

1. Introduction

Inner speech plays an important role in our understanding of human reasoning. We often hear our own voice “speaking” in natural language inside our own heads, especially while performing effortful cognition like planning and reasoning. Many people report experiencing a near-constant inner monologue that appears to track the subject of conscious attention and introspective awareness. We often talk through the considerations supporting one possibility or another when working through a problem, as though engaged in a dialogue or debate with ourselves. These self-narratives can help us solve complex problems by breaking them down into simpler steps, directing attention and reinforcing memory for important features, and allowing us to mentally explore different strategies for solving a problem before deciding upon a course of action. In fact, people who engage in more reflection and intermediate steps in inner speech when deliberating on difficult tasks tend to produce more accurate answers on difficult problems, such as multi-step arithmetic calculations and logic puzzles.

It may thus come as no surprise that similar benefits accrue to state-of-the-art artificial intelligence systems that “talk to themselves” before answering difficult problems. Specifically, a new generation of Large Language Models have been trained to produce internal “Chains-of-Thought” (CoT) before answering user queries. Self-prompting models have demonstrated remarkable improvement in performance on reasoning problems compared to previous versions of these architectures. These models—like OpenAI’s o1-o3, AnthropicAI’s Claude 3.7 Sonnet, and DeepSeek’s r1—distinguish themselves by engaging in long chains of textual self-prompting before producing an answer to a query. As with human inner speech, this text often appears to articulate inferential steps that led to their selected answers. These models are trained to self-prompt, often first through supervised learning on successful examples of intermediate reasoning steps obtained from humans when asked to “show their work”. Later, they are trained to adaptively produce CoTs through additional rounds of trial-and-error-based reinforcement learning, typically by giving them access to an algorithmic answer verifier that transmits a reinforcement signal as to whether the answer provided was correct or incorrect (called “Reinforcement Learning from Verifiable Rewards”, or RLVR). This trial-and-error learning allows models to discover which forms of self-talk are most effective in which contexts (Wen et al., 2025). Using these techniques, self-prompting models have achieved dramatic accuracy gains of up to 50% on difficult problem sets—such as the US Math Olympiad challenge, which requires subjects to not only solve novel and difficult mathematical problems but also to construct correct proofs (Riehl, 2025).

In a splashy press release, OpenAI claimed that this improvement in performance is evidence that these self-prompting models can “reason”—and indeed, “Large Reasoning Model” (LRM) has become the standard label for such models (OpenAI o1 contributors, 2024). OpenAI further claims that scrutinizing the CoTs these models produce can allow us to “read their minds”. These claims imply solutions to some of the most pressing philosophical and practical problems surrounding LLM-based artificial intelligence, especially the so-called “black box” problem regarding the opacity of their massively-complex internal operations. This opacity creates downstream practical considerations, making it more difficult to align model behavior with human values and to determine whether they can be deployed in a safe and trustworthy manner. Philosophers should thus be interested in understanding and evaluating such claims, not only because these practical questions have been paramount in the burgeoning philosophical work on the trustworthiness, safety, and value alignment of frontier AI systems, but also because the topics of inference, reasoning, and rationality have long been central to philosophy of mind, epistemology, and action theory. Just as previous AI advances created alluring new computational models with which we might explore classic ideas about perception, abstraction, and language, this new generation of systems may provide a conceptual playground to explore ideas about artificial reasoning.

Fortunately, inner speech’s role in human reasoning has been subjected to intense scrutiny in philosophy of psychology, mind, and action over the last twenty years. This work has also explored the opacity of human cognition and the question of whether inner speech reports of reasoning steps could reveal the structure of the rational transitions between thoughts. This literature has generally invoked a more demanding notion of reasoning than has been favored so far in research on LRMs in AI. Not every successful performance on a reasoning problem counts as reasoning; to jumpstart these intuitions at the outset, consider the difference between memorizing the answer to a familiar symbolic logic argument and actually working through intermediate steps by recognizing inferential support relations. Specifically, the richer notions require agents to construe their evidence as rationally supporting their conclusions in particular ways, and to draw their conclusions because of their sensitivity to these support relations. In addition to helping us predict and control the course of reasoning by understanding which structural features of problems drive reasoning outcomes, these construals could ground judgments of reasoning correctness or incorrectness, depending upon their alignment with norms of good reasoning. We might think that much of the excitement around LRMs and the purported ability of their CoTs to allow us to “read their minds” might be derived from this richer notion of reasoning—based on the idea that reading natural language CoTs could render the

decision-making of these massive and complex models rationally intelligible to us by providing access to the model’s individualized interpretation of its evidence.

Linking recent advances in LRMs to philosophical reflection on the nature of inner speech’s role in human reasoning should thus benefit both disciplines. As LRM-based systems become more capably agent-like, AI might draw increased inspiration from philosophical theories regarding the way that human agents are thought to control the course of their own reasoning through sensitivity to their evidence. Philosophy, in turn, could gain a new model system with which to articulate and explore, with unprecedented empirical detail and control, traditional claims about the ways in which inner speech and linguistically-vehicled reflection can support more robust, reliable, and intelligible reasoning. On the other hand, if the two discipline’s notions of “reasoning” cannot be aligned, this outcome might also be instructive—for we would have exposed and hopefully obviated the danger of overinterpretation and talking-past, and the behavioral success of LRMs on difficult reasoning problems might undermine inflationary theories of reasoning in human psychology.

In this paper, we begin Section 2 by reviewing self-prompting LRMs and their performance gains on reasoning problems, and in Section 3 we explicate three distinct ways in which the CoTs produced by LRMs might grant us a better understanding of their construals of their evidence. In Section 4, often by analogy with human cognition, we evaluate whether scrutiny of CoTs alone could provide these forms of understanding. We end with a somewhat pessimistic appraisal about the evidential value of CoTs considered in isolation from other forms of evidence, because—as is likely the case with human inner speech—tokens in CoTs can only causally influence a model’s answer choices by later being interpreted by the model (through self-attention). In Section 5, we will briefly consider whether supplementing CoTs with new techniques in mechanistic interpretability—which intervene on internal activations of models to reveal the causal roles that particular representations play in generating their behavior—might reveal the models’ internal construals of their evidence. While recent progress in mechanistic interpretability is promising and interesting, arbitrating fine-grained differences between distinct construals might not be resolvable by mechanistic interpretability techniques alone, but only when supplemented with a particular semantic theory of representation. We conclude in Section 6 by arguing that while there is a danger that philosophical interpretations of reasoning have been overintellectualized, engineers should not dismiss this research entirely—for practical guidance for improved self-calibration of chains-of-thought can be extracted from recent work in philosophy of psychology which interprets such construals as

epistemic feelings that reveal affordances for modulating reasoning behaviors in specific ways without requiring luminous access to internal representations.

2. “Large Reasoning Models” in Context

2.1 *A quick review of basic LRMs*

LRMs extend the basic LLM architecture—autoregressive transformers—by introducing components that are trained to adaptively engage in extensive (and often private) natural language self-prompting to enhance their performance on difficult multi-step reasoning and planning problems.¹ Until recently, however, pure transformer-based LLMs struggled with multi-step reasoning problems, demonstrating poor generalization performance on math, deduction, and planning puzzles that share the same underlying analytical structure as problems in the training set (e.g. the same logical form or underlying math equation) but differ in statistical structure (e.g. involving different predicates, attributes, examples, or difficulty levels). Reasoning has long been thought a central weakness of artificial neural network architectures, and many critics concluded that, as impressive as their performance gains have been on similarity and pattern-matching problems, LLMs are still not capable of genuine reasoning (Kambhampati, 2024; Mitchell, 2025). Many other components and complexities could be added to this brief review of LLM architecture and training, but these details will suffice to introduce the innovation of CoT self-prompting.

The idea to add self-prompting capability to the basic LLM architectures was inspired by the discovery that careful “prompt engineering” could improve their performance on difficult questions. For example, simply adding “Let’s think step by step” at the end of a human-provided prompt improved performance on simple arithmetic problems by up to 50% on earlier models like GPT-3 (Kojima et al., 2022). An even more effective intervention provides the model with a sample of human reasoning reports where intermediate steps are written out in text, which the model can then emulate before it provides its own answer. This led to a burgeoning study of effective prompt design, including training programs and guidebooks that purported to teach humans the best ways to pose their queries to LLMs. Much of this guidance is still relevant for state-of-the-art systems today. However, it has proven difficult to predict in advance which prompt modifications would be most effective, and researchers quickly realized that it might be more efficient to train the systems to discover the most efficient and effective ways to prompt themselves, instead of relying on humans to discover effective prompts through experimentation. The most impactful innovation behind LRMs

¹ For philosophically-oriented reviews of basic LLMs, see (Levinstein, 2023; Millière & Buckner, 2024).

probably involves their use of reinforcement learning on verifiable rewards (RLVR) to train models to adaptively self-prompt before answering difficult questions (Wen et al., 2025). Due to the cost of generating gold-standard answers on reasoning problems manually (using human labor), RLVR uses automated methods at scale to verify answers for correctness. RLVR has thus produced improvements on domains with objectively definable standards of success like mathematics and programming. In these contexts, mathematical algorithms or code compilers can be run separately to check an answer’s correctness.

Whatever their initial inspirations or similarity to human reasoning, it is uncontroversial that adding such self-prompting capability to standard LLM architectures enhances their performance on exactly the difficult reasoning and planning problems that most stymied earlier models. For example, on the 2024 US Math Olympiad, GPT-4o was only able to solve 12% of problems on average, supporting the position of skeptics that while LLMs may appear to demonstrate thinking and understanding, they are really just “stochastic parrots” that mimic textual concomitants of reasoning found in their training sets (OpenAI o1 contributors, 2024). In September 2024, however, OpenAI released its LRM-based “o1” model, which performed much better than GPT-4o on reasoning tasks—with performance gains of over 50% on several tasks. For example, o1 achieved 89th percentile performance on competitive programming questions, and it would have scored among the top 500 students in the US on the Math Olympiad qualifier, solving 74% of the difficult problems on average. Even more recent LRM-based versions of Gemini and o4 deploying extensive inference-time CoTs performed even better on the Math Olympiad, achieving Gold Medal performance. Technical results support these practical achievements by showing that LRMs can solve some problems that are in principle beyond the grasp of single-pass transformers that cannot self-generate CoTs (Z. Li et al., 2024; Yang et al., 2025).

The transcripts of CoTs released by OpenAI read like models reasoning in a way that reflects systematic sensitivity to the structure of their evidence. For a variety of challenging problems—cracking novel ciphers, solving physics problems, and engaging in tricky coding tasks—the CoTs appear to show o1 engaging in elaborate multi-step procedures that are linked to one another by relevant evidential relationships and systematically support the conclusions drawn. In the CoTs, the models seem to pursue the implications of various theories through speculation and deduction, monitor progress towards a solution by evaluating intermediate steps, backtrack to try out different hypotheses if the current track appears unpromising, and check the consistency of potential answers before terminating the CoT and delivering a final answer to the user. Of course, whether the natural

epistemic interpretation of this English text can be accepted at face value, however, is the central research question of this paper.

2.2 *Three Lines of Evidence Challenging the Transparency of CoTs*

The possibility that a model’s CoTs might “show the work” they performed to arrive at their answers has led to claims that these systems overcome the “black box” problems associated with deep neural networks. OpenAI, for instance, claims that we could use CoTs to “read the mind” of the model (*Learning to Reason With LLMs*, 2024)—further suggesting that this has obvious implications for safety and trustworthiness, because we might be able to search the CoTs for signs of unsafe behavior like deception. Some have claimed that perhaps CoTs could be used to learn how the model reasons when it produces a solution for any given output (Lyu et al., 2023). Philosophers have also recently taken an interest in the transparency-enhancing potential of CoTs; for example, David Chalmers (2025) has recently suggested that the CoTs produced by these models might be used to learn what kinds of attitudes the model has towards different propositions, perhaps marking an important step towards a goal in “propositional interpretability” that he calls “thought-logging”.

However, the claim that CoTs accurately represent the steps taken by a model to produce their answers has already drawn empirically-supported challenges. Let us consider several reasons which have already emerged to be skeptical about the accuracy of the CoTs as accounts of the way that LRMs arrived at their answers.

2.2.1 CoTs can omit causally relevant factors

Early experiments with CoT-producing models showed that they can systematically omit important causes of their answer responses. For example, Turpin et al. (2023) showed that models fine-tuned on biased input (e.g. on multiple choice tests where “C” is always the right answer) will reliably choose in ways consistent with the input bias (e.g. always choosing “C” as the correct answer, even when the text in “C” is clearly an incorrect answer to the question), but will fail to mention the biased preference in their justifications for their selection. Models in these experiments also frequently “rationalized” their choices in CoTs, but again without mentioning the choice bias. Granted, social psychology has found that humans readily concoct ad hoc rationalizations for choices to justify unwitting biases across a number of different settings (Cushman, 2018), from studies on priming to choice blindness to the “hungry judge” effect—where judges issue stricter parole rulings leading up to a lunch break, and more lenient rulings after the break, but without ever citing hunger or fatigue in their judicial decisions (Danziger et al., 2011). Sensitivity to extraneous factors not cited in justifications thus cannot disqualify models from parity with human reasoning,

but it does call into question the fidelity of the CoTs as representations of the underlying computational processes responsible for decision-making.

2.2.2 CoTs mention causally irrelevant factors

Relatedly, Lanham et al. (2023) attempted to assess the degree to which CoTs produced by Claude are epiphenomenal rationalizations, by either artificially truncating the CoTs at various arbitrary lengths or inserting mistakes in CoTs before allowing models to answer. If CoT tokens influence chosen answers by becoming part of the prompt context, these interventions on the CoTs should have influenced the model’s answer choices. Based on the difficulty and structure of the problems being solved, these researchers observed wide variation in the degree to which the modifications to the CoTs played a role in generating an answer (with the answer changing 10-60% of the time, depending on the kind of question being asked). When mistakes were inserted into CoTs, the researchers observed varying but modest degrees to which this altered the model’s answers, ranging from almost no change on questions thought to be “easy” for the models to solve without CoT, to 15% on questions thought to be more difficult. Both lines of evidence were taken to suggest that models largely select answers independently of the particular CoT tokens produced, and that the tokens selected may reflect epiphenomenal rationalizations—though again this may not reflect any fundamental difference with human cognition, as humans are also known to be habitual rationalizers (Cushman 2018).

2.2.3 The causal roles of CoT components may be opaque to human readers (steganography)

Another worry is that models may use CoTs to encode information steganographically, in a private manner that is unintelligible to human readers. McCarthy et al. (2025), for example, found that when forbidden from using standard symbols for mathematical operations (like “*” or “x” for multiplication), frontier models readily deploy alternative symbols to solve math and reasoning problems, with the meanings of these alternative symbols not obvious to naïve human monitors. Notably, however, these studies only found evidence of steganography after applying artificial constraints to the CoTs. Relatedly, other studies have found that training models to use “filler tokens” like ellipses or silent pause tokens before answering questions can improve performance on difficult reasoning problems, even though the tokens produced (e.g. a long series of ellipses) have no transparent meaning to human readers (Goyal et al., 2024; Pfau et al., 2024). The possibility of such hidden or encoded reasoning in CoTs has led to some general lessons that are often adopted by modelers, such that modelers should apply fine-tuning pressures to answers only (and not the CoTs

themselves, as this may introduce incentives to reason steganographically or covertly), and that models may interpret CoT tokens in ways that are different than human monitors.

3. Do CoTs “faithfully” report the LRM’s construal of its evidence?

Questions about the degree to which CoTs accurately portray the structure of underlying reasoning processes have been posed in computer science by asking whether CoTs are “faithful” representations of the model’s underlying reasoning. The notion of faithfulness here is proprietary, which we trace back to its origins in earlier methods of explainable AI.

3.1 The behaviorist character of existing faithfulness research

The relevant notion of faithfulness is often traced to an influential paper by Jacovi and Goldberg (2020), which reviews methods developed to explain the workings of pre-LRM neural network architectures. Jacovi and Goldberg’s treatment has several key features: 1) they distinguish an interpretation’s faithfulness from its plausibility (whether the proffered reasoning is satisfying or acceptable to humans), 2) insist that faithfulness comes in degrees, rather than being all-or-nothing and 3) lay out three common assumptions they diagnose in previous faithfulness research. These assumptions include: i) the “model assumption”, that two models make the same predictions if and only if they use the same reasoning process, ii) the “prediction assumption”, that faithful interpretations recommend the same decisions if and only if its reasoning is similar to that of the interpreted model, and iii) the “linearity assumption”, that some parts of a model’s input are more important to its reasoning than others, and the contributions of different parts of the reasoning are independent of one another.

Though this analysis inspired much subsequent work on faithfulness in the CoTs produced by LRMs, Jacovi & Goldberg’s criteria would likely be considered problematic by the standards of philosophical work on reasoning. For example, their model assumption is explicitly behaviorist in character; it would require identical interpretations of the reasoning of two systems whose decisions were input-output equivalent. However, it has long been a majority view in philosophy and cognitive science that two different reasoning processes could produce the same decisions but for different reasons or using different algorithms (Block, 1981; Marr, 1982). There are two straightforward arguments supporting this worry, one based on coextensive environmental features and another based on the general convergence of different methods which approach optimal inference. We review each in turn.

The first form of worry notes that features are often conflated in decision environments, and different processes can reach the same outcomes by processing different aspects or construals of the

same categories. A time-honored philosophical example here concerns “creature with a heart” and “creature with a kidney”; if all creatures with organs for pumping blood also have organs for filtering waste products from blood, then these two categories are *extensionally* equivalent, but *intensionally* distinct. They pick out the same creatures, but they do so using different characteristics; as a result, one could reason about creatures with hearts without necessarily reasoning in terms of creatures with kidneys, even though two decision processes based on the two features would end up making all the same categorization decisions. Jacovi and Goldberg’s model assumption would rule out the possibility that faithfulness research could reveal the same category being construed in two different ways. This example may sound like an abstruse philosophical quibble, but more workaday examples of conflated features raise pressing concerns in the literature on model safety and alignment.

For example, AI ethicists routinely worry about the problem of proxy variables, where models may discover conflated alternatives to protected variables, such as using zip code as a proxy for race. Conflated features are also routine concerns when evaluating model robustness and guarding against reward hacking, such as computer vision models that classify objects by texture rather than shape when those attributes are conflated in a training set, or hack a reward signal when the reward proxy (e.g. game score) is conflated with the true underlying objective (completing a level). Evaluations of robustness, trustworthiness, and safety fundamentally concern the generalization of a model’s computations from training environments that may conflate features to more open-ended testing environments where those conflations may come apart.² The philosophical contention that reasoning is not a purely extensional phenomenon might thus matter for practical evaluations in AI today; it matters to evaluations of robustness and trustworthiness how reasoning agents construe their evidence as supporting their decisions, rather than merely the attributes to which they have demonstrated causal sensitivity in the past. Thus, Jacovi & Goldberg should probably have only insisted on a weaker necessary condition for faithfulness: that interpretation methods should take models to use the same reasoning processes *only if* they make the same decisions in response to the same situations.

² The input-output equivalence in question is thus relative to some range of environments. If there is some environment where two models make different decisions, then there is some range of inputs where the two models are not input-output equivalent. However, because only a finite number of environments can be tested in interpretability or safety research, a high degree of input-output equivalence is enough to support the arguments of the remainder of the paper. Philosophers have also worried about intensionally distinct properties that are necessarily extensionally equivalent in all possible worlds (e.g. Quine’s “gavagai” problem—1960). However, we will not need to leverage coextensionality in all possible worlds to illustrate the practical problems here.

The second form of this worry notes that in cases with a single optimal solution to a problem, multiple methods might converge upon it but using different computational methods. Most relevantly to the present case, in many cases both Artificial Neural Networks (ANNs) trained by gradient descent and Bayesian models which adjust their priors to new evidence using varieties of Bayes’ rule are optimizing methods that asymptotically converge on the true probability distribution of a task given enough of a sample or learning (Cybenko, 1989; Freedman, 1963). However, the computations by which they achieve these results are only similar at a very high grain of abstraction; ANNs are generally not thought to literally compute Bayesian probability calculations, and Bayesian models are not learning the same representations as artificial neural networks. Like the prior form of the worry, they may deploy different construals of features in their intermediate computations that achieve the same ordinal rankings of options or comparable overall classification accuracy.

We might wonder here why machine learning has been tempted by behaviorist assumptions about the nature of reasoning, when cognitive science has long regarded reasoning processes to bear a many-to-one relationship with behavioral input-output profiles. The reason is probably practical: it has proven difficult or at least controversial to provide non-behaviorist methods to decode the internal representational structure of these massive, complex new models. A recent paper by AnthropicAI, for example, observes, in the course of adapting Jacovi & Goldberg’s criteria for faithfulness, that because of well-known black box problems, independent purchase on a model’s internal reasoning remains “a priori hard or impossible to determine”, even with complete access to its internal parameter values (Chen et al., 2025).

3.2 Moving faithfulness towards philosophical criteria for rational inference

In this section, we will evaluate the status of CoT’s as a list of reasons for the model’s inferences. Specifically, we provide an intensional criterion for faithfulness derived from philosophy of human reasoning, which is designed to highlight the oft-observed distinction between justifying an inference post hoc and rationally explaining it. To begin specifying this criterion, we can formally define a CoT C as a sequence of statements encoded in natural language, which is produced by a model in the course of arriving at answer O in response to query Q . The primary philosophical question we pose here is whether C could provide a faithful account of a model’s reasoning, understood in a richer intensional sense than has typically been invoked in the literature thus far on LRMs and faithfulness. We will find that this question can be explicated in several different ways that focus on different aspects of reasoning.

In evaluating CoTs for faithfulness as representations of the model’s reasoning, it will be useful to remember a common distinction between two different kinds of reasons they could provide for a model’s decisions:

(JR) **Justificatory reasons:** an argument from C *justifies* the choice of O (amongst alternatives) in response to Q , or

(ER) **Explanatory reasons:** an argument from C refers to the set of considerations, the recognition of which *actually caused* the system to produce O in response to Q .

It is important to see that these two types of reasons can be orthogonal; good reasons may justify without explaining (if they were not recognized in the reasoning that produced the conclusion), and bad reasons may explain without justifying (as with the hungry judge denying bail). The same set of arguments may be both justificatory and explanatory for O —perhaps this situation would routinely be the case for an ideally rational and calibrated reasoner—but justificatory reasons may be useful even if they are not explanatory. For example, they may provide good evidence for O ’s truth, or may establish precedents which are consistent with the previous choice and should be adhered to in the future as general policy. That said, we will set aside justificatory reasons for most of what follows; if faithfulness is to be useful in addressing the opacity, robustness, or trustworthiness of models, then it primarily concerns the degree to which C provides a set of explanatory reasons for the model’s production of O in response to Q .

The “recognition” of the reasons for producing O in response to Q in ER is often taken by philosophers to imply a necessary condition on rational inference that philosophers have called “Taking”. In an influential treatment, Paul Boghossian (2014) spells out the Taking condition as follows:

Taking: Inferring necessarily involves the thinker taking [their] premises to support [their] conclusion and drawing [their] conclusion because of that fact.

Some philosophers interpret Taking in a quite demanding way, for example as requiring conscious recognition of the inferential connections in an argument. We might thus wonder whether there is any real hope of satisfying Taking in the case of LLM-based agents, which may not be conscious in the relevant sense. Most philosophers have rejected such a demanding requirement for all instances of Taking, however, given regress arguments that would seem to make base cases of inference impossible in even humans. If all rational inferences required an explicit conscious recognition of the fit between premises and a conclusion it supports, then it would seem that we also need a conscious meta-recognition that we have assessed first-order evidential fit correctly, which would require an

even higher meta-recognition that the second-order fit was assessed correctly, and so on—leading to a vicious regress (Carroll, 1895; Fodor, 2000; Stroud, 1979). For these reasons, most philosophers in this area have argued that at least base cases of rational inferences must satisfy Taking in a way that does not require explicit meta-recognition. Philosophers have then focused on the kind of assessment between evidence and conclusion that should be performed for an inference to count as rational or reasons-responsive in the relevant way.

Perhaps the most popular and straightforward criteria has been rule-based—the idea that a thinker might simply possess a computational mechanism that can (automatically) assess whether premises support a conclusion by fitting their representational structure to the schema of a formal rule, such as from logic, probability, or decision-theory. Quilty-Dunn and Mandelbaum (Quilty-Dunn & Mandelbaum, 2018), for example, argue that to avoid regress, at least some of our rational inferences must be what they call Bare Inferential Transitions (BITs). BITs satisfy Taking simply by being constituted like a clause in a theorem prover algorithm; that is, they are causally sensitive to the propositional form of linguistically-vehicled assertions in such a way that the mechanism causes the agent to endorse the conclusion if it endorses the premises. For example, there may be one mechanism that causes BITs corresponding to Modus Ponens, where if the agent has evidence of the form “If P then Q” and also “P”, then it will Take those premises to support “Q”.

An alternative and less demanding interpretation of Taking requires only non-logical, content-based categorization judgments (Buckner, 2019; Shea, 2024). On this model, inferential connections depend on content-based assessments of features that can come apart at the compositional level—the way that we recognize someone’s face as a complex configuration of facial features arranged in a particular way. Though this model might initially sound less suitable as a candidate for rational inferences, it is actually quite powerful. For example, though Sherlock Holmes is often described as being a “master of deduction”, many of his most famous inferences are actually content-based inductions and abductions. In a famous train of thought from “The Sign of Four”, for example, Holmes astounds Watson by concluding that Watson had an older brother who was a drunkard, on the basis of inspecting Watson’s pocket watch. Holmes explains this chain of inference by articulating a series of categorization judgments: that the watch is not Watson’s because the initials on it do not match his given name but do match his surname, that valuable watches are usually inherited by older siblings, and that the watch was owned by a drunkard on the basis of scratches around its winding mechanism and dates etched on the inside from multiple pawnshop indices. These steps are all based on the arrangement of particular features possessed by the watch at

the exclusion of others, and matching them to those associated with a series of other categories—*valuable, inherited by older brother, and owned by a drunkard*. Holmes’ reasoning becomes intelligible when he lists these features and explains to Watson how he took them to support a sequence of intermediate supporting categorizations.

Such a list of featural assessments can provide explanatory reasons for DNN-based decision-making. For example, even though a certain object shape and a certain texture might be conflated in a wide range of training environments (e.g. as a wolf shape and wolf fur visual texture might be coincide in a training set for a computer vision model), they can be distinguished at the level of microfeatures that compose them—with one implicating causal sensitivity to lines and edges at the periphery of an object in an image, and another implicating contrasts and shadings distributed throughout the object’s area in the image. Only the features that were actually represented and assessed by the model can be truly said to satisfy ER and thereby explain the model’s answers.

3.3 Three intensional aspects of CoT faithfulness

With these clarifications in place, we now return to the central question: What might be revealed about the structure of an LRM’s reasoning (if it does anything worth the name) by reading its CoTs? The appeal to formal, rule-based criterion for Taking has a straightforward link to the fact that CoTs are linguistically-structured; perhaps the most ambitious position here is that CoTs provide a faithful record of formal inferences appreciated by the model in the course of producing the answer—as though reporting structural notes on its ongoing BITs. Call this criterion Syntactic Faithfulness:

Syntactic Faithfulness: The syntactic structure of the sentences provided in a CoT reveal the formal structure of the underlying computations that the model performed to convert the input provided in its prompt to the output it delivers as its final answer (and intermediate conclusions along the way).

This would provide us both with an account of the factors which led to the production of the model’s answer, as well as a formally adequate understanding of the computations the model performed on those factors to support the answer. The “takings” here would come in the form of formal syntactic patterns fit to the arguments provided in the prompt or intermediate steps in the CoT.

Given what we know of the internal operations of transformers, however, syntactic faithfulness seems implausible as a source of ERs for LRM-based inference, at least in frontier models. It is a routine discovery of robustness research that the deductive inferences of LRMs are subject to adversarial perturbations and generalize poorly to test samples far outside of the statistical

distribution of their training set samples (Krishna et al., 2025), which should not be the case if these models were fully competent in formal deduction rules. Such deviations between apparently rule-like justifications offered in a CoT and the model’s behavior and its choices in formally-identical but out of distribution test items would have to be written off as “performance errors” diverging from an underlying competence. Since it remains controversial whether transformers can robustly master symbolic logic even when that is their explicit training goal (Liu et al., 2025), perhaps we want to consider other senses in which a CoT could reveal an LRM’s reasoning.³ We will return to syntactic faithfulness in Section 5, however, when we consider mechanistic interpretability research which purports to show that at least some toy transformer models can exactly implement rule-like classical algorithms.

Alternatively, perhaps the inferential transitions within a CoT do not correspond to rule-based inferences, but rather to similarity-based categorizations like those in the Sherlock Holmes example reviewed above. The CoTs might report the features to which a model responds, and also suggest certain assessments as to whether categories listed elsewhere in a chain of reasoning “match” those features to a sufficient degree of abstract similarity. For example, if a CoT included the sentences “The platypus lives in water and has scales, but it also produces milk and is warm-blooded. Probably it is a mammal”, there is an apparent link drawn between “platypus” and “mammal”, and several intermediary features are suggested as supporting the link. Rather than the schema of a formal rule, perhaps this suggests something closer to the idea that the model possesses a stereotype or prototype for the category of mammal, and has determined based on some assessment of salient factors that the platypus adequately satisfies it. We may leave open the possibility that the assessment of features depends upon a context-sensitive arrangement, as we might with facial recognition (e.g. the interpretation of a nose or eyebrow may depend upon its location relative to other features) or the location of scratches on a watch (e.g. whether they are near the winding mechanism or on the watch face). Call the context-sensitive arrangement of features associated with a category a “configuration”, as is common in studies of configural learning in psychology. We can characterize this approach to satisfying Taking as Configural Faithfulness:

³ Researchers are exploring ways to tune CoTs to be more syntactically accurate by fine-tuning them on formal measures of coherence (Xia et al., 2025), though it is not clear if this merely improves their justificatory (rather than explanatory) status.

Configural Faithfulness: The CoT shows that the model has classified an entity E as a member of a particular category C , given a context-sensitive match between E ’s description and a configuration of features $F1-Fn$ it associated with C .

Compared to Syntactic Faithfulness, Configural Faithfulness may afford a better fit with previous hypotheses about the representational capacities of artificial neural network-based systems.

On this picture, arguments reported in CoTs correspond to a series of categorization judgments, revealing some of the underlying pattern-matching computations performed in selecting completions for chains (compared to alternatives).

This metric might be thought less demanding than Syntactic Faithfulness, but it would still articulate how CoTs might reveal the way underlying computations construed evidence as supporting a model’s answer to a query. This information satisfies some of the intuitions associated with Taking, which we argued above are relevant to robustness, safety, and trustworthiness. For example, a model satisfying Configural Faithfulness should mention in its CoTs texture cues like contrasts and shading if it is categorizing on the basis of texture, and shape cues like lines and angles if it is categorizing on the basis of shape (even if shape-based and texture-based decision-making recommend the same decisions in a dataset). Notably, Configural Faithfulness would still satisfy Taking in a sense, because different configurations could be associated with the same category. Configural Faithfulness should tell us which configuration of features was actually assessed in the pattern-matching judgment which produced the inference. Of course, the evidence for “gaps” reviewed in Section 2.2 suggests that models sometimes fail on even Configural Faithfulness, for example if they categorize on the basis of answer choice label but fail to report this in their CoTs. However, it is an intensional criterion that might be more easily satisfiable by DNN-based models because more of their inferences are likely to take this form.

Finally, there is qualitatively distinct kind of faithfulness we could assess at a macro-level as to how assessments of evidence influenced a model’s choice of answer: CoTs might reveal larger-scale reasoning strategies that a system pursues. For example, it might show when a model is backtracking to an earlier step in a reasoning process to try a different approach, breaking a problem down into subgoals, or rejecting a line of thought as fruitless. It might also show how options are ranked relative to one another in the final choice it makes before providing an output. We can dub this aspect of CoT evaluation “Strategic Faithfulness”:

Strategic Faithfulness: The CoT reveals relations amongst semi-independent arguments in a CoT that were causally-relevant to its selection of O in response to Q , such as backtracking, setting subgoals, and ranking candidate answers.

Of course, the arguments within a strategy will still need to be construed in representational terms, whether intensionally or extensionally; it will be difficult to describe the components of a reasoning strategy without providing some information about these representations that occur within it. But treated as black box components, qualitative options like backtracking and subgoal-setting might constitute independently evaluable dimensions of faithfulness. Strategic Faithfulness might provide some diagnostic and explanatory information unavailable from syntactic or configural faithfulness; for example, we might find that models provide more accurate answers on some types of problems if they backtrack more or less often, or that a particular string of tokens leading to a subgoal played a critical role in producing the answer. Strategic faithfulness thus constitutes a dimension which might be assessed and modulated independently of the others, for example by prompting a model to encourage more or less of certain higher-level strategies to improve performance (Gandhi et al., 2025).

4. Can a model’s CoTs satisfy intensional aspects of faithfulness?

With these criteria in place, we can now assess the degree to which CoTs provided by LRMs provide faithful evidence that they are in fact reasoning, according to intensional criteria unpacked from the Taking condition. To simplify the discussion in this section, we will focus only on the evidence that can be provided by CoTs alone, rather than CoTs in tandem or buttressed by additional sources of evidence, such as mechanistic analysis of internal model weights (which will be considered in Section 5). To complete this evaluation in a fair manner, we argue that it will be useful to compare the kind of evidence provided by CoTs produced by LLM-based agents to reports of inner speech in cases of human reasoning. A concern quickly revealed by this comparison is that philosophers and psychologists have often doubted whether inner speech provides transparent access to the thought processes underlying even human reasoning. Keeping this concern in mind will prevent us from overintellectualizing the evaluation of CoTs in LRMs, while also connecting the machine learning literature to mature discussions in psychology and philosophy regarding the roles that self-produced language may or may not play in human reasoning. We will conclude this section with a fairly pessimistic appraisal of the evidential values of CoTs themselves: in both LRMs and humans, the mere presence of an internal linguistic trace does not tell us much about the intensional construal of the agent’s evidence in underlying computations.

4.1 Prospects for syntactic faithfulness: the “blindness of introspection” challenge

Perhaps the strongest reading of the claim that CoTs allow us to “read the mind” of a model can be expressed as follows, aligned with ideal Syntactic Faithfulness:

- (a) We can read the mind of the model only if (1) the thoughts of the model are linguistic in nature and (2) syntactic relations in CoT track causally-efficacious propositional relations amongst the computations performed by the model to produce its answer.

After all, CoTs are linguistically-structured in text; if CoTs provide a transparent form of “thought logging”, then the inferential transitions between reasoning steps should respect syntactic relationships apparent in these CoTs, which in turn track their propositional relations via logical form. This would require that the underlying representations of the model are linguistic in format and that the inferential transitions between them track their syntactic structure.

A similar position on the relationship between inner speech and the underlying computations of human reasoning has been articulated in philosophy under the name of “sententialism”.

Sententialists hold that cognitive processes are constituted by representations that are themselves linguistic. Some sententialists argue that cognitive processes operate over a proprietary language that does not correspond to natural language, so-called *Mentalese* (Fodor, 1975, 2008; Jackendoff, 1989); in this case, giving a faithful account of reasoning in inner speech would involve translating accurately from *Mentalese*. Other sententialists argue that our cognitive processes are defined over natural language expressions, and that inner speech episodes can often directly reveal our thoughts (Carruthers, 1998). On this view, natural language expressed in inner speech might serve as the medium of thought itself. A weaker variant of this position might hold that inner speech episodes serve as a translation or paraphrase of more basic operations conducted in “*Mentalese*”, perhaps a translation or paraphrase which can in turn influence the course of later *Mentalese* inferences in various ways.

Sententialists often suggest that the linguistic format of inner speech episodes provides evidence for their view. Carruthers in particular has stated that “introspection informs us, in fact, that many of our thoughts are expressed in natural language” (1998, p50), though he has more recently come to adopt a hybrid sententialist position which holds there are both introspectible natural language thoughts and more basic language-like operations conducted non-consciously in lower-level systems in *Mentalese* (Carruthers, 2002). This line of argument—from the apparent format of inner speech episodes to the logical structure of our underlying thoughts and inferences—could mirror a similar move in AI today, from the linguistic structure of CoTs to a corresponding

propositional structure for the underlying thoughts and inferential transitions required for Syntactic Faithfulness.

In the human case, Machery (2005) argues that our inner speech episodes do not provide evidence that our thoughts are sentential (or linguistic) in nature. The argument unfolds as follows:

- A. Sententialism concerns the vehicles of our thoughts.
- B. We do not have any introspective access to the properties of the vehicles of our thoughts.
- C. Hence, the introspective fact of inner speech cannot be evidence that our conscious thoughts are linguistic.

Machery dubs this conclusion the “blindness of introspection” thesis. He grants that the apparently linguistic nature of our inner speech phenomenology does provide evidence that we represent the linguistic properties of these sentences in some sense, but it does not provide evidence that the representational vehicles over which our cognition is computed themselves possess these linguistic properties.

Machery supports this argument by analogy to conscious visual imagery. If we visualize a green triangle, then we know we are representing a green triangle; however, we know nothing about the syntactic properties of the representational vehicles over which this visualization is computed. In particular, we do not know that there is anything in our heads that is green or triangular. Similarly, when we experience linguistic auditory phenomenology in inner speech episodes, we know that we are representing certain linguistic properties of these natural language sentences; but we do not know that the representations over which this inner speech episode is computed themselves possess linguistic properties that could be exploited by syntax-sensitive inferential transitions. Downstream systems interpreting these episodes may be able to process these episodes in a language-like way, but only if they independently possess representations and inferential procedures with language-like structure. Imagine, by comparison, that a telepathy experiment suddenly granted a dog access to the auditory phenomenology corresponding to its owner’s inner speech; the dog would be able to exploit the propositional information embedded in the syntactic structure of this auditory stream only to the same degree that it could already understand such structure in the master’s spoken linguistic commands.

In the case of LRMs, we do have natural language sentences printed in an internal CoT scratchpad, but this does not guarantee that the LRM can take advantage of this structure in syntax-sensitive formal reasoning operations. Once the linguistic structure is present in an internal CoT, it becomes part of an LRM’s unfolding prompt. This means that downstream self-attention operations

have access to that structure in drawing their conclusions. However, like the dog granted access to a human’s inner speech, if these self-attention operations are not already enabling computations that respect classical formal syntax, adding representations with this structure to the prompt will not help. The CoTs themselves provide no privileged form of evidence that underlying operations are sensitive to syntactic inferential transitions—as opposed to merely mimicking inferences consistent with those transitions in contexts similar to those in its training set, in the way a struggling logic student might mimic argument forms on familiar examples from memory without appreciating the inferential support relations behind them.

This is not to say that it is impossible that the underlying computations of LRMs are sensitive to syntactic structure in the right way; only that CoTs that apparently demonstrate this do not provide evidence that they are. In other words, the success of syntactic faithfulness would stand or fall with independent evidence of the ability of transformers to implement classical algorithmic computations on syntactic structures present in their prompt windows. We will return to those sources of evidence in section 5.

4.2 Prospects for configural faithfulness: Feature representations and mechanistic interpretability techniques

What, then, of the prospects for configural faithfulness in the CoTs generated by LRMs? Indeed, we might think these prospects more promising; whereas formal syntactic computation of the sort required by BITs has long been thought a distinctive weakness of ANN architectures, configural pattern-matching has long been seen as a distinctive strength. Many theorists have supposed that category representations in ANNs are more like feature-based prototypes that fit this kind of configural representation very well (Clark, 1993; Way, 1997) and some of the most successful models of configural category learning are artificial neural networks (Gluck & Myers, 1993). Outputting words corresponding to features of a category might also be thought to make those features available for categorization via self-attention, by appending them to the prompt which is used to predict future token outputs. We might expect this to cue up a configural take for later inferential processing by underlying computational mechanisms.

However, there are already good empirical reasons to doubt the link between the construal of a category in a model’s CoT’s or human inner speech and its actual influence in underlying computations. Again, there seems to be a symmetry: the text of the inner speech may portray the construal in a rosier way than its actual influence, either for social and psychological benefits of

positive self-conception (in the human case) or due to reward-hacking of fine-tuning and safety training signals (in the LLM case).

In the human case, philosophers and psychologists have doubted that inner speech in humans could provide transparent access to the semantic contents of our own mental states. One cluster of positions here has been called “inferentialism”, which holds that inner speech only provides us with an additional source of input which must be interpreted, and from which we can *infer* the intensional contents of our mental states, in the same way we infer the mental states of others from their statements and behavior.⁴ Regardless of whether this position is correct in the human case, it should be the default position in the LLM case, because we know from architectural details that CoT tokens are generated in the same way as a model’s “overt” speech—using autoregressive next-token prediction driven by self-attention—and that CoT tokens can only influence later processing through the same mechanism that the overt speech of others affects their later inferences, through being decoded, as tokens appended to the prompt, by later self-attention operations. So, whether inferentialists are correct about the human case, it seems they must be correct about LRMs as a matter of known architecture.

Thus, if we are considering the evidence that the CoT itself provides for an intensional construal of the features the system took to support a categorization judgment, we will face problems similar to those we faced with syntactic faithfulness. The occurrence of a name for a feature in a CoT suggests that a feature is available for self-attention operations to drive categorization decisions under the hood, but it does not establish how this information was interpreted by the system or even whether the information was in fact used by the system at all. The generic reasons for skepticism about the faithfulness of CoTs reviewed in Section 2.2 amplify concerns that the model may interpret tokens differently than their conventional meanings to human observers. Moreover, in the case of conflated proxy variables, it does not establish which of a word or phrase’s extensionally equivalent interpretations has been applied by the model. Indeed, conventional fine-tuning and alignment training regimes may exacerbate the distance between the text produced in a CoT and underlying computations driving decisions. For example, a system might have been fine-tuned to avoid mentioning protected characteristics like race, so it instead mentions a set of features that provide a proxy for race in its CoTs, which in turn activate race representations in the underlying computations actually driving its decisions when interpreted by self-attention. If we

⁴ For details of such a view, the most prominent contemporary defence of inferentialism is probably due to Peter Carruthers (2009, 2011)

were hoping that CoTs would help us rule out importantly distinct intensions that are extensionally-conflated in an environment, then, we will be left with few reasons for increased confidence. According to reports from Apollo Research (2025), Claude 3.7 Sonnet has even appeared to track hints that its CoTs are being evaluated for faithfulness in a testing scenario, providing more circumspect CoTs in those situations.

Returning to the literature on the role of inner speech in human cognition, some theorists respond to inferentialists by holding that inner speech statements provide a special kind of forward-looking access to our mental states by committing us to adhere to articulated policies more systematically in the future. In this sense, stating a policy or justification for something in inner speech may help us “make up our minds” between different construals of reasons for acting. This is an interesting idea to explore in the context of LRMs, especially based on the idea that different “thought branches” appear to pursue different sets of assumptions to their conclusions before models select the “best” answer. However, even state-of-the-art LRMs are thought to be systematically unstable in just the ways we might hope such a forward-looking commitment view might help in the human case. The CoTs of even state-of-the-art LRMs still show more incoherence and instability than humans, with more dithering between options and ignorance of pursuing inconsistencies across longer stretches of the prompt window and on more complex problems (Shojaee et al., 2025). Reducing forms of reasoning instability like “sycophancy” and “inconsistency” remain important research targets for these groups. Notably, however, there have been steady improvements along these dimension in the latest models, especially as evidenced by data reported in the Claude 4.5 Sonnet system card. Current models still likely exhibit substantial limitations on the forms of agential self-determination that might help achieve that forward-directed consistency, but it may be that models in the near future will make impressive enough gains in this direction to count as “making up their minds” in this self-directed way.

To summarize, where does this leave us with respect to the evidence that CoTs might provide for configural faithfulness? Perhaps in a somewhat stronger default position than with respect to syntactic faithfulness, but a position that still remains quite weak and unable to systematically rebut concerns about intensional inaccuracy relevant to important issues like the problem of proxy variables. We know by default that ANNs like LRMs excel at representing context-sensitive feature representations required for configural takes in categorization judgments. However, the CoTs produced by LRMs can only influence such takes by being text tokens appended to the prompt which must in turn be interpreted through self-attention. As in the human case, there are

reasons to suppose that the interpretation of these tokens might be systematically distorted through fine-tuning to misrepresent the actual interpretations under the hood. There remains an interesting possibility that stating certain features as reasons for a categorization may commit a model to more consistently categorize in that way in the future, perhaps even committing them to a self-interpretation of previously ambiguous statements. However, the fact that models “reset” their prompt windows between sessions and still struggle with sycophancy and consistency concerns suggest that there is not yet a human-level ability to stabilize future behavior using such self-avowals.

4.3 Prospects for strategic faithfulness

Strategic faithfulness might be thought more promising than the other two aspects of faithfulness, because it depends less on fine-grained discriminations between distinct intensional interpretations of underlying computations or representations. We may observe, for example, a system appearing to backtrack in a CoT to an earlier strategy by favoring the answer it seemed to suppose that strategy favored, whether or not we can characterize the links in that CoT with much intensional accuracy. Getting in the right intensional neighborhood may be enough. If the CoT is even approximately accurate, this approach will offer at least have some contrastive utility: while we might not be able to distinguish Strategy 1 from Strategy 1’ which is identical except cast in terms of extensionally-conflated proxy variables, we will typically be able to distinguish Strategy 1 from a different Strategy 2 if they favor different answers, however the intermediate links in the chain are construed.

A skeptic here may worry that strategic faithfulness is parasitic on other notions of faithfulness that allow us to characterize the components of a strategy in an intensionally-satisfying way. Consider the space of strategic behaviors explored by Gandhi et al. (2025): verification, subgoal setting, backtracking, and backward chaining. Again, the only evidence considered here is an apparent statement of the subgoals in the text of a CoT; to what degree can we determine which subgoals were set without an ability to characterize those subgoals in an intensionally-determinate way and compare them to the subgoals stated in the CoT? If we turn our attention to verification, how can we tell whether an answer-choice option was appropriately verified, without knowing whether the computation underlying the verification corresponds to the option stated in the CoT?

Nevertheless, we can, as Gandhi et al. show, modify LRM behavior to display more or less of certain strategic behaviors without settling these questions. We can prime models with examples of CoTs that display certain strategic behaviors, causing them to produce more future CoTs that are consistent with these various reasoning strategies. We can then investigate relationships between

exhibiting more or less of these behaviors and improved or diminished accuracy on a specific type of reasoning problem. This kind of control would certainly be useful for practical purposes relevant to reasoning performance, and should not be dismissed. It would not, however, satisfy the intensional demands typically unpacked from Taking. Without an ability to characterize reasoning steps in intensional terms, it might not be possible in difficult cases to say whether an earlier strategy to which the model apparently backtracked was really distinct from a later one it abandoned, or whether an apparent inconsistency exposed by checking for incoherence was really contradictory. With respect to intensional criteria about the way agents are construing their evidence, then, we might see strategic faithfulness as an importantly distinct notion which can be studied with different methods, and one which has its own kind of practical importance for improving outcomes on reasoning problems, but which is fundamentally derivative on syntactic or configural faithfulness to type the reasoning steps which occur in the strategies it characterizes.

5. Can mechanistic interpretability save intensional faithfulness?

Standard methods of interpreting deep neural networks over the last ten years have often included probing methods which show that parts of a neural network’s activation space tracks information about target categories and attributes which they might use for categorization or problem-solving (Ravichander et al. 2021). Linear probes—discovery methods that show simple combinations of activation values from network components can predict other attributes in the environment—have often shown that information tracking the status of grammatical properties of sentences or features of items relevant to category decisions can be predicted from network activations with high accuracy. We might hope that these techniques would help us determine the faithfulness of an LRM’s CoT on some range of reported categorization decisions. We could, for example, direct linear probes towards features mentioned in a model’s CoTs to see whether that information is really present in network activations, and consider a positive probing outcome as evidence that the CoT is faithful to underlying reasoning processes. The mere presence of such information, however, does not establish that the network actually uses this information to drive its output categorizations or decisions (Harding, 2023; Millière & Buckner, 2024). These networks and their inputs are often large and highly complex, and information present in a network may merely reflect perturbation effects from complex input with minimal consequences on the network’s computations. Eliminating or disrupting information in internal activations, for example, sometimes has little effect on network performance on the tasks (Elazar et al., 2021).

For this reason, researchers have looked to philosophy of science in search of stronger methods for explaining the causal role information may play in these networks. There is now a vast array of “interventionist” or “mechanistic” interpretability methods which purport to show evidence of a semantically-interpretable causal structure in these networks using interventionist approaches to causal explanation from philosophy of science (Geiger et al., 2021; Woodward, 2003). These methods begin by using linear probes to identify regions of activation space for hypothesis-driven interventions on network activations, and then attempt to confirm or disconfirm that the information plays the hypothesized role in generating network outputs by modifying that information in a systematic way. We will discuss briefly one of the most philosophically-grounded such methods, the causal abstraction framework from (Geiger et al., 2021, 2025).

In very rough terms, the causal abstraction framework proposes causal graphs structured around interpretable features that a network may use to compute its categorizations. It then tests an alignment between this interpretable causal graph and the trained network’s performance by intervening on parts of the network’s activations that probes have identified as potentially implementing those parts of the causal graph.⁵ The simplest form of intervention is an ablation, which simply disrupts the information in the network by disabling nodes or disrupting activations until that information is no longer present in the network’s activity. Ablation is however one relatively crude technique in a family of what these theorists call “interchange interventions”, many of which can be used to alter the information in networks in more natural, hypothesis driven ways, as though we were swapping out the valence of a represented entity’s property with another value that it realizes on another input. For example, we can attempt to toggle the system’s representation of a noun from singular to plural, or the network’s representation of a container from full to empty, by patching part of the activation values carrying information about the entity’s current state with its representation of another entity’s state on another set of inputs (B. Z. Li et al., 2021; Ravfogel et al., 2021). If the network then changes its behavior in the hypothesized direction—it completes a sentence as though the noun was plural rather than singular, or as though the container was empty rather than full—this is taken to provide evidence that the information in the network plays a computational role as hypothesized by the causal graph to which it was aligned (Geiger et al. 2025). Causal abstraction might be thought to provide the missing link between CoTs and underlying

⁵ There are many complex constraints on interpretable interventions; for example, Harding (2023) articulates a “control” condition which requires that the intervention not disrupt the Kullback-Leibler divergence for other labels (a measure of the entropy in the probability distribution for other category representations). This is, of course, a strong idealization, but we do not explore these complexities here.

computations required for intensional forms of faithfulness. We could, for example, treat a LRM’s CoTs as sources of causal abstractions to test through interchange interventions—as we might with a causal graph depicting the features that Sherlock Holmes purports to use to categorize Watson’s watch as belonging to a drunk—and suppose that positive results from interchange experiments show that the CoT satisfies the desiderata of Configural Faithfulness.

Much more could be said about these fascinating and impressive new techniques for uncovering the causal organization of complex ANNs; but for present purposes, the important point to note is that they are unlikely to resolve questions about intensional characterization necessary to satisfy Taking. There are two related problems with relying on causal abstraction techniques to settle an LRM’s interpretations of its own evidence.

First, the results of activation patching experiments from the causal abstraction framework rarely find perfect and univocal alignment with a single interpretable causal graph. Perfect alignment can sometimes be found on small, toy examples, but it is rarely present in larger networks trained on more diverse data; that we should not expect perfect alignment for frontier models is consistent with findings that even the most powerful LRMs face complete performance collapse on formal problem structures above a certain complexity (Shojaee et al. 2025). Experiments instead reveal a range of accuracy rates at which the tested interventions are consistent with the computations performed over features in a causal graph. Different and conceptually incompatible causal graphs may at the same time present us with slightly different but still high rates of interchange intervention accuracy. Thus, in cases where we are unsure which configuration of features is used to categorize an input as being a member of a target category in an inference chain—such as cases of proxy variables that are conflated in an environment’s informational structure, mentioned above—interchange interventions will only reveal different degrees of support for different causal graphs. Simply assuming that the graph with the highest tested intervention accuracy depicts the true category intension assessed by the network seems premature; indeed, assessing a different set of interchange interventions could provide reversed comparative accuracy rates. Instead of concluding that the network implements the algorithm with the highest intervention accuracy for some set of interventions, it may be more appropriate to conclude that it implements a potentially inconsistent superposition or hodgepodge of different algorithms that it toggles between in an ad hoc manner.

Second and relatedly, causal abstraction relies on a prior interpretation of task structure to interpret the results of interventions. Specifically, justifying the representational interpretation of an intervention relies on what Harding (2023) calls the “Misrepresentation” condition, as the possibility

of misrepresentation has long been considered an essential feature of informational approaches to representation. According to the traditional story, semantic interpretations of causal components are unnecessary if it is not possible for them to say something incorrect about the world; this is what distinguishes a mere “causal relay” (which merely transmits information from current environmental circumstances) from a representation (Dretske, 1986; Ramsey, 2007). According to Harding, a component obtained by causal abstraction satisfies the Misrepresentation condition if and only if applying a “correct” intervention to an activation vector—toggling it from storing an incorrect value of some variable to a correct value, according to the causal abstraction—improves outputs in an expected way. “Correct” here is a technical term defined with respect to what the modeler regards as “the true label distribution” for some task. In other words, causal abstractions are defined relative to a modeler’s independent construal of success on a task. This seems to exclude the possibility that causal abstraction could reveal the model’s own subjective interpretation of its evidence if its interpretation of the task structure differs from what the modeler regards as correct.

Such divergences, however, appear in exactly the sorts of cases where we would expect mechanistic interpretability to settle interpretational disagreements. For example, suppose that a certain texture and a certain shape are conflated in a model’s training environment, as has commonly been observed for computer vision models. If the modeler regards the shape category as the correct label for this feature, causal abstraction could not determine that the model interprets the evidence in terms of texture, because responding to texture and not shape in post-training environments where the two diverge would be regarded by the modeler as a mistake which could be corrected by intervention (in other words, the model could be found to misrepresent shape as texture, but not be found to misrepresent texture as shape). This would rule out by fiat a possibility allowed by other metasemantic theories, that the model’s interpretation of its evidence differs in normative structure from our own interpretation of its evidence—exactly the possibility we would not want to close off when considering intensional approaches to interpretability.

Different theorists will offer different diagnoses for what has gone wrong here, perhaps depending on their theory of representation. Non-naturalists may suppose that the lack of a conscious first-person perspective on the world implies that there is no fact of the matter as to how the networks take their inputs to support their outputs (Mendelovici, 2020). More naturalistic approaches may suppose that the normative interpretation of a causal graph should be settled by aspects of a network’s history which go beyond the scope of interchange interventions, such as those derived from etiologial representational functions. For example, teleosemantic theorists may

suppose that the issue is settled by informational relationships which held in the past during a privileged period of selection (Dretske, 1988; Mollo & Millière, 2023), though features which were informationally conflated during the training history or fell well short of perfect informational relationships will still result in indeterminate representational attributions (Slater, 1994).

What is really needed to determine intensional accuracy for Taking, we might think, would be some sense of which of the candidate interpretations of the network’s causal structure is favored *from the artificial agent’s own perspective*, without closing off the possibility that it interprets the task in a different way than we do. However, it is unclear how to assess this question using modifications of current mechanistic interpretability techniques. Of course, LRMs can produce text, and could be prompted (or self-prompted) to consider different interpretations of their own current policies. This raises the alluring possibility that we may be able to settle such questions for LRMs by simply asking them. Combined with a tendency to exhibit future behavior that is more consistent with what they have stated in the past, this may play a role similar to “cashing out” intuitions that some non-naturalist philosophers think settle such intensional questions in the human case (Mendelovici, 2013, 2020). With humans, however, we want to know not just which interpretation we favor in a momentary judgment, but rather from a *considered* perspective, one towards which our judgments stabilize given adequate time and resources for consideration. It is unclear whether LRMs are yet sufficiently sophisticated, stable, or embedded in a rich social context for it to interpret as having such a considered perspective—a possibility that their well-documented susceptibility towards sycophancy and adversarial attack still speak against. As a result, while mechanistic interpretability techniques may give us an impressive degree of control over a system’s performance, control which is consistent with a variety of mutually-inconsistent algorithmic characterizations, it is unlikely to give us the kind of univocal interpretation of category intensions required for Taking.

Given the detour through mechanistic interpretability, it may help to summarize the concerns of this section at a higher level of description. The intensional aspects of faithfulness pertain to the particular combinations of features (syntactic, configural, or strategic) of the agent’s evidence that it takes to support its conclusion. For a CoT to provide evidence of the features which the agent takes to support its conclusions, we would need to determine the correct interpretation of the agent’s causal organization which produced that decision, which mechanistic interpretability purports to do. However, mechanistic interpretability methods only provide us with a range of possible representational construals of the agent’s causal organization which are consistent with intervention experiments to varying degrees. Because agents can make mistakes, behave inconsistently, or endorse

an interpretation of the normative task structure that differs from our own, the varying degrees of consistency with different algorithmic models obtained from mechanistic interpretability studies provides the wrong kind of evidence to decide which algorithm captures the agent’s own perspective on its evidence. While these various causal models of the structure of the agent’s assessment may give us some real information about its current causal organization that is useful for purposes like control and prediction, without the addition of a metasemantic theory, they cannot themselves settle doubts about univocal takes sufficient for intensional faithfulness.

6. Does it matter for AI research if LRMs cannot satisfy intensional notions of faithfulness?

Computer scientists and more pragmatically-minded philosophers may feel impatient at this point, wondering whether it matters to pressing practical goals of LRM development and evaluation if our interpretability methods cannot satisfy intensional notions of faithfulness. The arguments of the previous section turned on semantic and normative issues that may appear abstracted and distant from engineering goals involving control over LRM behavior. This may be coupled with a sense that intensional criteria for rational inference are overintellectualized in even the human case.

In this final section, we temper this dissatisfaction with evidence that internal signals which are sensitive to intensional construals of evidence can play important roles in rapid metacognitive calibration of inferences in humans, animals, and potentially AIs. These forms of control possess clear practical importance both for improving agent architectures in LRMs, providing additional layers of self- and operator-control over their reasoning trajectories. To rebut the sense that intensional notions of faithfulness are over-intellectualized, we will take a short detour through the research on multiple decision making systems in earlier AI architectures and in the metacognition of nonlinguistic animals, highlighting the ways that such signals may be specialized into a variety of epistemic feelings in linguistic, culturally-scaffolded humans. In situations with which we have had significant practice and experience, we may be able to perform well using almost automated skill, with little sensitivity to our assessments of our evidence from one moment to the next. However, such decision procedures may operate inefficiently or at chance in novel situations with which we lack the extensive experience required for such skilled automation. In those cases, internally-generated, intensionally-sensitive metacognitive signals may allow us to more rapidly calibrate inferences to new situations.

To begin with an analogy from the distant AI past, in 2015 Researchers at DeepMind discovered the benefit of multiple decision pathways early in the deep learning revolution, from their experiences training systems via reinforcement learning to play Atari games (Blundell et al., 2016).

Given thousands of hours of gameplay experience, these systems could learn to play most games at superhuman levels of skill. This performance dropped to chance or worse, however, in situations where reinforcement learning had not provided extensive opportunities to optimize gameplay policies to environmental contingencies. In these cases, a much simpler decision rule prevailed: maintain an explicit database of recent situations, moves, and their outcomes, and at the next iteration simply do whatever action previously produced the highest reward outcome in the most similar situation to the one currently encountered. DeepMind researchers called the architecture which allowed this short-term database lookup option the “episodic controller architecture”, on analogy to the way that humans might consult episodic memory for rapid decision-making. A system with an episodic controller has two decision pathways it could consult to solve any problem in a gameplay frame: either apply an optimal policy learned from extensive prior experience, or, where this policy failed to produce a high confidence outcome, simply retrieve a successful action from the most similar recent episode. Whereas the standard reinforcement learning regimes might require tens or hundreds of thousands of repetitions to discover an optimal policy, the episodic controller architecture could often find acceptable moves after being exposed to similar situations just a few times.

While this example pertains to episodic-like recollection in reinforcement learning agents, a structurally similar case has been made for the importance of internally-generated metacognitive signals in research on inference in nonlinguistic animals. Animal cognition research provides an excellent comparison case to determine whether intensional criteria like ‘Taking have been overintellectualized. In the case of animals, extensive evidence has been collected in “uncertainty monitoring” experiments which purport to show that animals are sensitive to the degree to which their current perceptual evidence supports their decisions. A common design involves asking animals to categorize a stimulus as a member of one of two categories, where some range of stimuli in the test set are near the threshold where animals cannot reliably discriminate members of one category from another. Dot-pattern designs are commonly used here; animals are trained to respond in one way if there are few dots and another way if there are many dots, with some range of dot-patterns in an intermediate range which are difficult for animals to discriminate. A correct categorization earns the animal an immediate reward (e.g. a macaque monkey may receive a grape), and an incorrect response is punished with a long timeout before the subject is provided with a new trial. Animals are then given a third, “bailout” response option which provides no reward but immediately provides them with a new trial. In this design, many different species of animal learned to adaptively use the

bailout option on stimuli in the intermediate range where they had low categorization accuracy, providing *prima facie* evidence that they had some internal signal providing them with information about the strength of their current evidence.

Skeptics of animal metacognition were not convinced by this design, however, for they argued that adaptive use of the bailout could be explained by its associative reward history without appealing to special metacognitive signals. The bailout option, these skeptics argued, acquired its own associative reinforcement history by being indirectly associated with rewards and timeouts. In other words, animals might not need any kind of internal signal providing them with information about the strength of their current evidence to use the bailout option adaptively; given enough experience, they could simply use the option in cases where the dot-pattern stimulus was sufficiently similar to cases where the use of the bailout option made them more likely to receive a grape on the next trial more quickly, rather than after a long timeout.

To overcome this skeptical interpretation, Smith et al. (2006, 2008) devised a clever modification of the bailout design, where they delivered rewards and timeouts not as trial-by-trial feedback after each decision, but only in aggregate at the end of a large block of trials. This should prevent animals from receiving direct associative feedback as to whether they were using the bailout option adaptively through reinforcement learning. Despite this absence of feedback signals, some subjects—especially a particularly adept macaque named “Murph”—nevertheless learned to use the bailout option adaptively on intermediate threshold trials where his discrimination accuracy dropped. Notably, Murph’s use of the bailout option did not perfectly track his objective discrimination accuracy, but appeared to show a subjective bias towards some trials on one side of the threshold, including some range of trials where his discrimination accuracy was high. This suggests that not only did Murph possess some internal metacognitive signal providing information about the strength of current perceptual evidence for a decision, but also that this signal reflected a subjective assessment which did not perfectly track objective decision success. For Murph, this signal might reflect something subjectively generated by the psychophysics of the discrimination context, which Murph was able to harness, through something like executive control, to adaptively select the bailout option.

By hypothesis, such an internal subjective assessment of inferential strength could play an important role in current LRMs the same sorts of situations where DeepMind decided they needed to add an episodic lookup table to their Atari-playing agent—those where systems lack extensive training feedback on the outcomes of their inferences. In the base case, we may suppose that such

signals would be experienced as a sense of uncertainty when assessing whether a particular set of evidence supports a particular decision. Such a sense of uncertainty is perhaps only one of a host of “epistemic feelings” which can provide a subjective assessment of evidential qualities relevant to a mental action, which can be used for decision-making. Other epistemic feelings include tip-of-the-tongue phenomena (for memory retrieval), feeling of knowing (regarding a potential answer to a query), and feeling of forgetting (that one has omitted or neglected something related to current tasks). In humans, though these signals are fallible and (like Murph’s metacognitive signal) potentially biased, they routinely help guide mental actions like inferring, deciding, remembering, and task-switching.

Several theorists have recently proposed that such epistemic feelings may function like mental affordances that, in response to some tacit sensitivity to internal task-relevant representations, highlight opportunities to regulate and calibrate mental actions (Arango-Muñoz, 2014; McClelland, 2019, 2020). To perceive an affordance is to perceive an opportunity for action—in this case, particular mental actions. For example, the tip-of-the-tongue phenomenon affords us additional memory search in nearby locations to attempt to retrieve a memory or fact. In the case of inferences and CoTs, the hope is instead that internal signals afford opportunities to affirm conclusions (whether interim or final), re-assess evidence, collect alternative evidence, explore an alternative strategy, or discard an inferential policy as unpromising. Crucially, they may do so in a way that is sensitive to intensional construals of evidence—particular combinations of features assessed, subjective strength, or alternative policies—without needing to grant luminous access to the structure of those features or policies. For example, two inferences drawing upon different sets of features that recommend extensionally equivalent conclusions might generate different epistemic feelings and thereby different affordances for inferential calibration, without supposing that the agent need have explicit, discursive access to the sources of those feelings. Just as some theorists have argued that epistemic feelings offer an attractive way to finesse the intellectual demands of Taking in the case of humans and animals (Buckner 2020, Shea 2024), functionally similar internal signals may play a similar role in AI agents (McClelland, 2017).

This line of thought may be particularly encouraged by recent findings concerning epistemic transition points in CoTs—so-called “aha” or “inflection” moments, which empirical evidence suggests are some of the most faithful components of CoTs in frontier LRMs. These transition points are operationalized in CoTs by tracking interjective phrases such as “wait”, “actually”, “no, that’s wrong”, “I just realized...”, “aha!”, or “Oh, I see now”. As noted in earlier sections, decoding

studies increasingly find that models appear to have committed themselves to answers before outputting CoT tokens that provide evidence leading to those conclusions, suggesting that much of reasoning traces are epiphenomenal rationalizations (Feng et al., 2024; Wang, 2026). Nevertheless, researchers have also found that tokens corresponding to epistemic transition markers occur almost exclusively in cases where probes reveal shifts in confidence between competing answers through the course of a CoT (Boppana et al., 2026). By tracking internal signals corresponding to these interjections, mechanistic interpretability researchers were able to decode an “early stopping” signal that predicted confidence in an answer well before the models themselves would terminate a CoT and express confidence in an answer choice. Decoding this internal confidence signal thus provided an additional layer of external control that researchers could use to calibrate CoT production and reasoning behavior, one which even generalizes to tasks for which models lacked extensive training. Such early exiting could save up to 80% of computation used on reasoning budgets, leading to substantial efficiency benefits without a loss of decision performance (Boppana et al. 2026). There is thus already strong evidence that internally-generated metacognitive signals of the sort described above are computed by models and can be used by both models and researchers to adaptively calibrate reasoning performance, at least with respect to gross answer confidence.

The open question that will likely be addressed by near-term research concerns whether these basic internal signals can be calibrated to recapitulate a full arsenal of more sophisticated epistemic feelings simply by learning on cultural materials already available in training sets, or whether more specialized epistemic feelings would require specialized training or architectural tweaks. We could, as in the human case, imagine a more sophisticated set of intensionally-sensitive epistemic feelings which could be used to calibrate inferences in low-data contexts, such as curiosity, boredom, feelings of knowing, feelings of coherence or integration with background knowledge, confidence in inferences to the best explanation, or the awe of creative insight. Researchers are already investigating the computation of multiple epistemic signals; Alizadeh et al. (2026), for example, leverage three separate sources of intrinsic metacognitive signal: self-consistency, reasoning trace length, and model self-verbalized confidence in inferential transitions. At present, however, these signals are still computed indirectly from decision statistics or model outputs—for example, self-consistency is operationalized as the number of CoT paths that converge on the same interim conclusions, and verbalized confidence simply asks the model itself to report on the strength of its evidential assessment—rather than from internal signals reflecting takes on evidence.

The next generation of developments in CoT faithfulness research may thus instead leverage internal activations and architectural modifications to detect and cultivate a set of epistemic feelings corresponding to epistemic relations that can be used for more sophisticated self- and operator-calibration of reasoning behaviors. Such feelings might look like three-place relations, corresponding to i) an epistemic attitude (such as curiosity, hesitancy, or awe, with corresponding functional roles), ii) a limited base of evidence marking the source of intensional fit, and iii) a conclusion potentially supported by that limited evidence base. We should then develop models which can leverage such emotion-like roles to develop their own distinctive epistemic perspective on occurrent inferences, deploying reasoning personalities which cannot be reduced to statistics obtained from training histories or external behaviors. This new generation of models will “wonder” whether a certain basis of evidence supports a particular conclusion, be “awed” by particularly creative inferences that integrate complex and difficult forms of evidence, and “doubt” conclusions arising from inferences which lack subjective support. Though it is not yet possible to predict whether this next generation of models deploying their own distinctive epistemic perspectives will require new architectural tweaks or whether sophisticated epistemic feelings can be scaffolded from cultural materials already available in training sets, extensions of existing mechanistic interpretability techniques and current training regimes appear adequate to legislate the question.

7. Conclusion

In this paper, we reviewed evidence that the adaptive use of CoTs in LRMs can improve their performance on difficult reasoning problems, as well as recent literature assessing the “faithfulness” of Cot text to underlying computations. Ambitious claims by development firms like OpenAI that scrutiny of CoTs would solve problems in interpretability and safety research by allowing us to “read the minds” of model are likely overstated and premature. Moreover, review of philosophical work on rational inference showed that such claims are likely to be misinterpreted as allowing us to understand the model’s own subjective epistemic takes on its evidence, whereas existing faithfulness research in computer science has not even attempted to address these questions. This would leave us with the conclusion that, while the production of CoTs by LRMs undoubtedly leads to performance gains on difficult reasoning problems, by the standards of philosophy and psychology of reasoning, we have little evidence as to whether LRMs are actually doing any real reasoning.

On the other hand, this conclusion could lead us to worry that intensional criteria for rational inference have been overintellectualized, a charge that has sometimes arisen as an internal critique in

philosophy and psychology themselves. Nevertheless, recent work in these areas has suggested ways to finesse demanding criteria like intensional Taking by reflecting on the nature of epistemic feelings, which might be constituted by internal signals which are sensitive to intensional construals of evidence which are sufficient for metacognitive calibration, but without granting luminous access to those grounds. This metacognitive calibration might provide important practical utility to models and modelers precisely in cases where current frontier models still struggle, in contexts where they have little training experience with problems of that sort. Evidence suggests that in such unfamiliar decision scenarios, humans and animals rely upon internally generated epistemic feelings to guide the collection and appraisal of evidence for inferences. We thus conclude with a suggestion to reframe the relevance of faithfulness and intensionality to the development of flexible reasoning behavior in AI agents. Instead of attempting to “read the minds” of models, we should focus on imbuing agents with epistemic feelings by scaffolding from cultural materials available in training sets or introducing architectural modifications corresponding to internally generated signals not reducible to training history and observable decision statistics.

References

- Alizadeh, K., Shojaei, P., Cho, M., & Farajtabar, M. (2026, March 7). *Recursive Language Models Meet Uncertainty: The Surprising Effectiveness of Self-Reflective Program Search for Long Context*. arXiv.Org. <https://arxiv.org/abs/2603.15653v1>
- Arango-Muñoz, S. (2014). Metacognitive Feelings, Self-Ascriptions and Mental Actions. *Philosophical Inquiries*, 2(1), 145–162. <https://doi.org/10.4454/philing.v2i1.81>
- Block, N. (1981). Psychologism and Behaviorism. *Philosophical Review*, 90(1), 5. <https://doi.org/10.2307/2184371>
- Blundell, C., Uria, B., Pritzel, A., Li, Y., Ruderman, A., Leibo, J. Z., Rae, J., Wierstra, D., & Hassabis, D. (2016). Model-free episodic control. *arXiv Preprint arXiv:1606.04460*.
- Boghossian, P. (2014). What is inference? *Philosophical Studies*, 169(1), 1–18. <https://doi.org/10.1007/s11098-012-9903-x>

- Boppana, S., Ma, A., Loeffler, M., Sarfati, R., Bigelow, E., Geiger, A., Lewis, O., & Merullo, J. (2026). Reasoning theater: Disentangling model beliefs from chain-of-thought. *arXiv Preprint arXiv:2603.05488*. <https://arxiv.org/abs/2603.05488>
- Buckner, C. (2019). Rational Inference: The Lowest Bounds. *Philosophy and Phenomenological Research*, 98(3), 697–724.
- Carroll, L. (1895). What the tortoise said to Achilles. *Mind*, 4(14), 278–280.
- Carruthers, P. (1998). *Language, thought and consciousness: An essay in philosophical psychology*. Cambridge University Press.
https://books.google.com/books?hl=en&lr=lang_en&id=bTQjfCm8tKwC&oi=fnd&pg=PR11&dq=carruthers+1996+natural+language&ots=aS2JYj8NwH&sig=p_jbFI69eh_SiaOpKcHbXhKgaQI
- Carruthers, P. (2002). The cognitive functions of language. *The Behavioral and Brain Sciences*, 25(6), 657–674; discussion 674–725.
- Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, 32(2), 121–138.
- Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. OUP Oxford.
- Chalmers, D. J. (2025, January 27). *Propositional Interpretability in Artificial Intelligence*. arXiv.Org.
<https://arxiv.org/abs/2501.15740v1>
- Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., & Perez, E. (2025). *Reasoning Models Don't Always Say What They Think* (arXiv:2505.05410). arXiv.
<https://doi.org/10.48550/arXiv.2505.05410>
- Clark, A. (1993). *Associative engines: Connectionism, concepts, and representational change*. MIT Press.
https://books.google.com/books?hl=en&lr=lang_en&id=G4fBLKAh-

- IUC&oi=fnd&pg=PR9&dq=prototypes+and+connectionist+networks+a+natural+fit&ots=856xHhOJFD&sig=PWqLmiVm4Yh0FgtRZ-m6zA3ZjFw
- Claude Sonnet 3.7 (often) knows when it's in alignment evaluations. (2025, March 17). *Apollo Research*. <https://www.apolloresearch.ai/blog/claude-sonnet-37-often-knows-when-its-in-alignment-evaluations/>
- Cushman, F. (2018). Rationalization is rational. *Behavioral and Brain Sciences*, 1–69.
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4), 303–314.
- Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17), 6889–6892.
<https://doi.org/10.1073/pnas.1018033108>
- Dretske, F. (1986). Misrepresentation. In R. Bogan (Ed.), *Belief: Form, content and function* (pp. 17–36). Oxford. http://books.google.com/books?hl=en&lr=lang_en&id=mv2vGDT-6KIC&oi=fnd&pg=PA297&dq=misrepresentation&ots=4DhbShB65C&sig=-6I4skrdbTQvssUUPg8SmYZVMxg
- Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. The MIT Press.
- Elazar, Y., Ravfogel, S., Jacovi, A., & Goldberg, Y. (2021). Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9, 160–175.
- Feng, J., Russell, S., & Steinhardt, J. (2024). Monitoring latent world states in language models with propositional probes. *arXiv Preprint arXiv:2406.19501*. <https://arxiv.org/abs/2406.19501>
- Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.
- Fodor, J. A. (2000). *The mind doesn't work that way: The scope and limits of computational psychology*. MIT press.

- Fodor, J. A. (2008). *LOT 2: The language of thought revisited*. Clarendon Press.
- http://books.google.com/books?id=0U_vi285S_cC&pgis=1
- Freedman, D. A. (1963). On the Asymptotic Behavior of Bayes' Estimates in the Discrete Case. *The Annals of Mathematical Statistics*, 34(4), 1386–1403.
- <https://doi.org/10.1214/aoms/1177703871>
- Gandhi, K., Chakravarthy, A., Singh, A., Lile, N., & Goodman, N. D. (2025). *Cognitive Behaviors that Enable Self-Improving Reasoners, or, Four Habits of Highly Effective STaRs* (arXiv:2503.01307). arXiv. <https://doi.org/10.48550/arXiv.2503.01307>
- Geiger, A., Harding, J., & Icard, T. (2025). *How Causal Abstraction Underpins Computational Explanation* (arXiv:2508.11214). arXiv. <https://doi.org/10.48550/arXiv.2508.11214>
- Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34, 9574–9586.
- Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, 3(4), 491–516.
- Goyal, S., Ji, Z., Rawat, A. S., Menon, A. K., Kumar, S., & Nagarajan, V. (2024). *Think before you speak: Training Language Models With Pause Tokens* (arXiv:2310.02226). arXiv. <http://arxiv.org/abs/2310.02226>
- Harding, J. (2023). Operationalising Representation in Natural Language Processing. *The British Journal for the Philosophy of Science*, 728685. <https://doi.org/10.1086/728685>
- Jackendoff, R. (1989). What is a Concept, that a Person May Grasp It?¹. *Mind & Language*, 4(1–2), 68–102. <https://doi.org/10.1111/j.1468-0017.1989.tb00243.x>
- Jacovi, A., & Goldberg, Y. (2020). *Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?* (arXiv:2004.03685). arXiv. <https://doi.org/10.48550/arXiv.2004.03685>

- Kambhampati, S. (2024). Can large language models reason and plan? *Annals of the New York Academy of Sciences*, 1534(1), 15–18. <https://doi.org/10.1111/nyas.15125>
- Kojima, T., Gu, S. (Shane), Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, 35, 22199–22213.
- Krishna, A., Rastogi, A., & Galinkin, E. (2025). *Weakest Link in the Chain: Security Vulnerabilities in Advanced Reasoning Models* (arXiv:2506.13726). arXiv. <https://doi.org/10.48550/arXiv.2506.13726>
- Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., ... Perez, E. (2023). *Measuring Faithfulness in Chain-of-Thought Reasoning* (arXiv:2307.13702). arXiv. <http://arxiv.org/abs/2307.13702>
- Levinstein, B. (2023, January 30). A Conceptual Guide to Transformers: Part I [Substack newsletter]. *The End of Reckoning*. <https://benlevinstein.substack.com/p/a-conceptual-guide-to-transformers>
- Li, B. Z., Nye, M., & Andreas, J. (2021). Implicit representations of meaning in neural language models. *arXiv Preprint arXiv:2106.00737*.
- Li, Z., Liu, H., Zhou, D., & Ma, T. (2024). Chain of thought empowers transformers to solve inherently serial problems. *arXiv Preprint arXiv:2402.12875*, 1. <https://www.academia.edu/download/115029747/2402.12875v2.pdf>
- Liu, H., Fu, Z., Ding, M., Ning, R., Zhang, C., Liu, X., & Zhang, Y. (2025). *Logical Reasoning in Large Language Models: A Survey* (arXiv:2502.09100). arXiv. <https://doi.org/10.48550/arXiv.2502.09100>
- Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., Apidianaki, M., & Callison-Burch, C. (2023). Faithful chain-of-thought reasoning. *Proceedings of the 13th International Joint Conference on*

- Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 305–329.
- <https://aclanthology.org/2023.ijcnlp-main.20.pdf>
- Machery, E. (2005). You don't know how you think: Introspection and language of thought. *The British Journal for the Philosophy of Science*, 56(3), 469–485.
- Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. *Henry Holt and Co., New York*.
- <http://library.wur.nl/WebQuery/clc/172778>
- McCarthy, R., Joey, S., Ibanez-Lissen, L., Georgiev, V., Watts, C., Whittingham, H., Gonzalez-Manzano, L., Tice, C., Young, E. J., & Radmard, P. (2025). Large language models can learn and generalize steganographic chain-of-thought under process supervision. *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*.
- <https://openreview.net/forum?id=2g5cJqX15Y>
- McClelland, T. (2017). AI and affordances for mental action. *Proceedings of the AISB Annual Convention 2017*, 372–379. <https://wrap.warwick.ac.uk/id/eprint/87246/>
- McClelland, T. (2019). Representing our options: The perception of affordances for bodily and mental action. *Journal of Consciousness Studies*, 26(3–4), 155–180.
- McClelland, T. (2020). The mental affordance hypothesis. *Mind*, 129(514), 401–427.
- Mendelovici, A. (2013). Reliable misrepresentation and tracking theories of mental representation. *Philosophical Studies*, 165(2), 421–443.
- Mendelovici, A. (2020). Propositional Attitudes as Self-Ascriptions. In *Common Sense Metaphysics*. Routledge.
- Millière, R., & Buckner, C. (2024). *A Philosophical Introduction to Language Models -- Part I: Continuity With Classic Debates* (arXiv:2401.03910). arXiv. <https://doi.org/10.48550/arXiv.2401.03910>

- Mitchell, M. (2025). Artificial intelligence learns to reason. *Science*, 387(6740), eadw5211.
<https://doi.org/10.1126/science.adw5211>
- Mollo, D. C., & Millière, R. (2023). *The Vector Grounding Problem* (arXiv:2304.01481). arXiv.
<https://doi.org/10.48550/arXiv.2304.01481>
- OpenAI o1 contributors. (2024). *Learning to reason with LLMs*. OpenAI.
<https://openai.com/index/learning-to-reason-with-llms/>
- Pfau, J., Merrill, W., & Bowman, S. R. (2024). *Let's Think Dot by Dot: Hidden Computation in Transformer Language Models* (arXiv:2404.15758). arXiv. <http://arxiv.org/abs/2404.15758>
- Quilty-Dunn, J., & Mandelbaum, E. (2018). Inferential Transitions. *Australasian Journal of Philosophy*, 96(3), 532–547. <https://doi.org/10.1080/00048402.2017.1358754>
- Quine, W. V. O. (1960). *Word and object*. MIT Press.
<http://books.google.com/books?id=sVH3YIIZK5QC&pgis=1>
<http://books.google.com/books?id=2OApYgEACAAJ&pgis=1>
- Ramsey, W. M. (2007). *Representation reconsidered*. Cambridge University Press.
- Ravfogel, S., Prasad, G., Linzen, T., & Goldberg, Y. (2021). Counterfactual Interventions Reveal the Causal Effect of Relative Clause Representations on Agreement Prediction. *arXiv Preprint arXiv:2105.06965*.
- Riehl, E. (2025, November 1). *AI Crushed the Math Olympiad—Or Did It?* Scientific American.
<https://www.scientificamerican.com/article/mathematicians-question-ai-performance-at-international-math-olympiad/>
- Shea, N. (2024). Metacognition of inferential transitions. *Journal of Philosophy*, 121(11).
<https://ora.ox.ac.uk/objects/uuid:1675a658-7d98-4838-9997-5936fc4ba963>
- Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens

- of Problem Complexity. *SuperIntelligence-Robotics-Safety & Alignment*, 2(6). <https://s-rsa.com/index.php/agi/article/view/15919>
- Slater, C. (1994). Discrimination Without Indication: Why Dretske Can't Lean on Learning. *Mind & Language*, 9(2), 163–180. <https://doi.org/10.1111/j.1468-0017.1994.tb00221.x>
- Smith, J. D., Beran, M. J., Couchman, J. J., & Coutinho, M. V. (2008). The comparative study of metacognition: Sharper paradigms, safer inferences. *Psychonomic Bulletin & Review*, 15(4), 679–691.
- Smith, J. D., Beran, M. J., Redford, J. S., & Washburn, D. A. (2006). Dissociating uncertainty responses and reinforcement signals in the comparative study of uncertainty monitoring. *Journal of Experimental Psychology: General*, 135(2), 282.
- Stroud, B. (1979). Inference, belief, and understanding. *Mind*, 88(1), 179–196.
- Turpin, M., Michael, J., Perez, E., & Bowman, S. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 74952–74965.
- Wang, W. (2026). LLM Reasoning Is Latent, Not the Chain of Thought. *arXiv Preprint arXiv:2604.15726*. <https://arxiv.org/abs/2604.15726>
- Way, E. C. (1997). Connectionism and Conceptual Structure. *American Behavioral Scientist*, 40(6), 729–753. <https://doi.org/10.1177/0002764297040006005>
- Wen, X., Liu, Z., Zheng, S., Ye, S., Wu, Z., Wang, Y., Xu, Z., Liang, X., Li, J., Miao, Z., Bian, J., & Yang, M. (2025). *Reinforcement Learning with Verifiable Rewards Implicitly Incentivizes Correct Reasoning in Base LLMs* (arXiv:2506.14245). arXiv. <https://doi.org/10.48550/arXiv.2506.14245>
- Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press. <http://books.google.com/books?id=LrAbrrj5te8C&pgis=1>

Xia, S., Li, X., Liu, Y., Wu, T., & Liu, P. (2025). Evaluating mathematical reasoning beyond accuracy.

Proceedings of the AAAI Conference on Artificial Intelligence, 39(26), 27723–27730.

<https://ojs.aaai.org/index.php/AAAI/article/view/34987>

Yang, C., Li, Z., & Wipf, D. (2025). Chain-of-thought provably enables learning the (otherwise)

unlearnable. *The Thirteenth International Conference on Learning Representations*.

<https://openreview.net/forum?id=N6pbLYLeej>