Understanding Is Not a Scalar: What the Chinese Room Could Not Imagine

Chang-Eop Kim
Gachon University
eopchang@gmail.com

Abstract

Searle's Chinese Room argument derives its persuasive force from a specific implementation image: a person following a static rulebook of symbol-manipulation rules. The "stochastic parrots" critique revived the same intuition for contemporary systems. We argue that this line of argument constitutes an *implementation-dependent* thought experiment—one whose intuitive force diminishes when its implementation premises change. Two premises underlying the original image—static symbolic manipulation and internal opacity—are violated in significant respects by modern learning systems. Models trained on different modalities and domains converge toward shared representational geometries, and within individual models, non-spatial inputs give rise to geometric representations that mirror known structures of the domain. Large language models exhibit concept-selective self-reports causally grounded in internal activation states. These findings do not prove understanding exists, but establish that the system image presupposed by the Chinese Room no longer describes how modern systems operate. We propose a multi-dimensional framework that decomposes understanding into three conceptually separable dimensions: structural (capturing world structure in representational geometry), self-modeling (causally grounded access to internal states), and phenomenal (subjective experience). Under this framework, the Chinese Room's conclusion is valid for systems matching its implementation profile—high phenomenal experience but near-zero structural capture and self-modeling—yet this profile no longer applies to modern learning systems, and the "stochastic parrot" characterization is revealed to conflate absence on one dimension with absence on all. The question becomes which dimensions of understanding contemporary learning systems instantiate, and which remain genuinely open.

1 Introduction

Searle's Chinese Room argument [Searle, 1980], published in 1980, remains the most influential philosophical critique of machine understanding. A person inside a room follows a rulebook of Chinese symbol-manipulation instructions to produce outputs indistinguishable from those of a native speaker, yet that person understands no Chinese—so the argument goes. From this thought experiment, Searle concluded that executing a program is never sufficient for understanding.

The argument has generated over four decades of debate. The systems reply (the room as a whole understands), the robot reply (sensory grounding would enable understanding), and the brain-simulator reply (replicating the brain's causal architecture would suffice) have all been advanced and rebutted in turn [Searle, 1980, 1990]; for a comprehensive survey, see Preston & Bishop [2002]. More recently, the "stochastic parrots" critique has revived essentially the same intuition for contemporary systems: large language models manipulate linguistic forms without access to meaning [Bender et al., 2021]. Borg [2025] argues that LLM outputs possess genuine semantic content through a process of semantic deference—belonging to an existing meaningful communicative practice—yet distinguishes this from original intentionality—illustrating how the debate has been refined rather than resolved. No consensus has been reached.

We diagnose this impasse as arising partly from two structural problems.

First, the Chinese Room is an *implementation-dependent thought experiment*—one whose intuitive force depends on a specific image of how the target system works: a person flipping through a static rulebook of discrete symbol-manipulation rules. The AI of 1980 did indeed work that way, but contemporary learning systems have systematically departed from those implementation premises. When the premises shift, the intuitions built upon them require re-examination.

Second, the debate has been conducted under the tacit

assumption that "understanding" is a single, indivisible property. A system either understands or it does not. The stochastic-parrots critique inherits this same binary framing—language models either access meaning or they do not [Bender et al., 2021]. Yet this all-or-nothing framing is inadequate for capturing the complex profile of contemporary learning systems—systems that form representations sensitive to world structure, that exhibit limited but causally grounded self-modeling, and whose phenomenal status remains indeterminate.

This paper makes two contributions. First, we introduce the concept of *implementation-dependence* for thought experiments and use it to analyze why the Chinese Room's intuitive force weakens for modern systems. To this end, we examine how two lines of recent empirical evidence—representational convergence and cross-domain structural transfer (Section 3.1), and causally grounded functional introspection (Section 3.2)—violate the implementation premises of the Chinese Room. Second, we propose a multi-dimensional framework that reconceptualizes understanding along three dimensions: structural, self-modeling, and phenomenal.

Our goal is not to refute the Chinese Room argument, nor to claim that modern AI systems "understand." Rather, it is to reformulate the question that the Chinese Room posed—"What constitutes understanding?"—into a more productive form. The central thesis is that understanding is not a scalar—not a single quantity that is either present or absent—but a multi-dimensional property whose components can be separately instantiated, investigated, and debated. Transforming the binary question "Does this system understand?" into the dimensional question "Along which dimensions of understanding, and to what degree, does this system operate?"—this is the step we propose.

2 The Chinese Room as an Implementation-Dependent Thought Experiment

Thought experiments are powerful tools for mobilizing intuitions in philosophy. However, not all thought experiments operate in the same way. We propose a distinction between two types—*structure-dependent* and *implementation-dependent* thought experiments—and argue that this distinction illuminates why the Chinese Room, specifically, requires re-examination in light of modern AI systems.

Structure-dependent thought experiments derive their intuitive force from the logical, epistemic, or moral structure of the situation itself. The trolley problem is a paradigmatic example: the moral tension between five lives and one persists whether the trolley is replaced by an autonomous vehicle or the specific technology changes entirely. The *validity scope* of such thought experiments is largely unaffected by changes in technological implementation.

Implementation-dependent thought experiments, by contrast, derive their intuitive force from a specific image of *how* a system operates. In this type, when the underlying implementation changes, the persuasiveness of the intuition changes with it. This distinction is better understood as a spectrum than a strict dichotomy. Some thought experiments (e.g., the trolley problem, the teleporter) derive their force almost entirely from logical structure; others (e.g., Jackson's Mary's Room; Jackson, 1982) occupy an intermediate position where implementation details amplify but do not constitute the core intuitive engine. The Chinese Room, we argue, falls toward the implementation-dependent end of this spectrum, and it is this positioning that makes it uniquely vulnerable to changes in AI technology.

Two complementary perspectives from the epistemology of thought experiments help clarify what is at stake. Norton [2004] argues that thought experiments are "disguised arguments"—their conclusions follow from premises, and their vivid narrative details ("picturesque particulars") serve only to make the underlying argument accessible. On Norton's account, if the empirical premises of a thought experiment change, the conclusion no longer follows automatically. This analysis directly supports our argument in Section 3: the Chinese Room's central assumption—that AI systems operate by following explicitly authored rules for manipulating discrete symbols—was empirically well-grounded in 1980 but is challenged by the operation of modern learning systems. By Norton's own logic, when this assumption loses its empirical basis, the thought experiment's conclusion requires independent re-justification.

However, Norton's framework leaves a puzzle unexplained: *why has the Chinese Room remained persuasive long after its premises became questionable?* If the thought experiment were purely a logical argument, its force should diminish as its premises weaken. Gendler [2004] offers an account that stands in tension with Norton's. Against Norton's reduction of thought experiments to arguments, Gendler argues that thought experiments also function as what we may call *cognitive simulations*—they generate "quasi-sensory intuitions" through imaginative engagement with a scenario, and these intuitions play an irreducible epistemic role that cannot be captured by premise-to-conclusion reasoning alone. We propose that the Chinese Room's persistence is at least partly explained by this mechanism. The vivid image of a person turning pages in a room functions as a cognitive simulation that generates intuitions *directly*—before the audience undertakes any premise-by-premise analysis. It is this simulation-driven intuitive force, rather than the formal structure of the underlying argument, that has sustained the Chinese

Room's influence across four decades of technological change. Our concept of implementation-dependence draws on *both* accounts despite the tension between them: from Norton, the insight that changed premises weaken conclusions; from Gendler, the insight that vivid imagery sustains intuitions independently of premises. It is precisely because thought experiments can operate on both levels simultaneously that implementation changes may destabilize the argument (Norton's level) while the cognitive simulation (Gendler's level) persists—or vice versa. An implementation-dependent thought experiment is one whose cognitive simulation is tightly bound to a specific image of how the target system operates, such that changes in the underlying technology destabilize both the argument's premises and the simulation's intuitive grip.

The Chinese Room is a paradigmatic case. Its persuasive power flows from a concrete image: a person inside a room turning pages in a rulebook, looking up which output symbols correspond to input symbols. The moment we imagine this image, our intuition reacts strongly: "This person does not understand Chinese." Searle's argument rides this intuitive reaction toward the conclusion that "merely executing a program is never sufficient for understanding." Block's "China Brain" thought experiment [Block, 1978]—in which an entire nation's population collectively simulates a brain's functional organization—provides another illustration. Block's own argument targets the logical possibility of absent qualia under functionalism, a point that does not depend on any particular implementation image. However, the *intuitive force* of the thought experiment—what makes the absence of consciousness seem obvious rather than merely logically possible—does depend on the concrete image of a billion people communicating via two-way radios. When the implementation image shifts (say, to a silicon chip replicating the same functional organization at nanosecond timescales), the logical argument remains intact, but the intuitions often shift. It is this latter, psychological phenomenon that our concept of implementation-dependence is designed to capture.

However, for the Chinese Room's cognitive simulation to generate its characteristic intuition, at least two implementation premises must hold in the background. These premises are the concrete form in which the thought experiment's central assumption—that AI operates by explicit rule-following over discrete symbols—was instantiated in 1980:

1. **The static-symbolic premise.** The system's operating rules are explicitly authored by an external designer and fixed prior to deployment. The symbols the system manipulates are individual and discrete; relations between symbols are limited to those explicitly stated in the rulebook, with no implicit similarity structure, distance relations, or clusters forming among them. The rulebook in the room is pre-written, never updated during use, and the symbols it manipulates remain atomic—"mountain" and "river" bear no systematic relationship within it. This absence of learned internal structure reinforces the intuition that the system has no access to the meaning of what it processes.

2. **The internal-opacity premise.** The system has no access to what kind of processing it is performing: neither forming internal states that are sensitive to content categories, nor producing self-reports that are causally connected to such states. Whether it is handling a question about emotions or geography makes no difference to the person inside the room.

These two premises were a reasonably accurate description of AI as it existed in 1980—rule-based expert systems and Good Old-Fashioned AI (GOFAI). Those systems did indeed operate by following human-authored rules to manipulate discrete symbols, with no learning-induced internal structure.

Contemporary learning systems, however, systematically depart from both premises. This is not merely a matter of technological progress; it matters because the Chinese Room's intuitive force—both as argument (Norton) and as cognitive simulation (Gendler)—is built precisely upon these premises. When the implementation premises collapse, the intuitions erected upon them require re-examination.

An important clarification is in order. Our claim is not that the Chinese Room's conclusion is false—that executing a program guarantees understanding. Rather, the claim is more precise: the *intuitive grounds* that made the conclusion compelling were built upon a specific implementation image, and that image no longer describes the systems to which the conclusion is most frequently applied. The argument's formal structure may be preserved, but its central assumption—that AI systems operate by explicit, externally authored symbol manipulation—now requires independent justification for systems whose operation bears little resemblance to the rulebook in Searle's room.

The next section examines how modern learning systems specifically depart from each of these premises, drawing on recent empirical research.

3 What Modern Learning Systems Have Changed

In Section 2, we analyzed the Chinese Room's intuitive persuasiveness as depending on two implementation premises: static symbolic manipulation and internal opacity. This section examines, through recent empirical work, how

modern learning systems violate each of these premises. A growing body of mechanistic interpretability research has documented evidence that LLMs form structured internal representations that track environmental regularities, and perform structured reasoning over them; for a comprehensive survey, see Beckmann & Queloz [2025]. Rather than surveying this literature exhaustively, we focus on two lines of empirical evidence that map onto the Chinese Room's implementation premises: representational convergence and cross-domain structural transfer (Section 3.1), which challenge the static-symbolic premise, and functional introspection (Section 3.2), which challenges the internal-opacity premise. Our goal is not to argue that "understanding has been demonstrated" but to show that the system image presupposed by the Chinese Room is fundamentally at odds with how modern systems actually operate. Where we introduce formal notation below, it is not intended as a mathematical reduction of understanding but as a precision tool to clarify what each line of evidence does and does not establish.

3.1 Violation of the Static-Symbolic Premise: Learned Representations and World Structure

The Chinese Room's rulebook is externally authored and unchanged during use. The person inside applies rules but never creates or modifies them. The symbols manipulated are discrete and individual; no implicit similarity structure or distance relations form among them. Modern neural networks operate in a qualitatively different manner. Parameters are shaped through learning from data, and the result is not a collection of explicit rules over atomic symbols but geometric structure in high-dimensional vector spaces—that is, *representations*—in which formerly discrete tokens acquire systematic relational structure.

Huh et al. [2024] report a striking phenomenon in this process, which they term the Platonic Representation Hypothesis. Models trained on different modalities (vision and language) develop increasingly similar internal representational geometries as their scale and performance grow. This convergence is observed through representational similarity measures such as kernel alignment and mutual nearest-neighbor analysis, and the authors interpret it as evidence that models are converging toward a shared statistical structure of reality. This finding builds on a growing body of evidence: Kornblith et al. [2019] demonstrated that networks of the same architecture trained from different initializations learn similar representations measurable via centered kernel alignment, Moschella et al. [2023] showed that latent spaces across diverse architectures and modalities preserve relative distance structure sufficiently for zero-shot model stitching, and Bansal et al. [2021] found that representations from supervised and self-supervised training regimes can be stitched together with minimal performance loss. Representational convergence, in short, is not an isolated curiosity but a robust and repeatedly observed phenomenon. This convergence extends beyond vision-language models: Edamadaka et al. [2025] compared scientific foundation models trained on different physical domains—small molecules, inorganic materials, and proteins—with different input formats and often non-overlapping training data, and found that their representational structures converge both within and across domains as performance improves, suggesting that they capture structural regularities of matter shared across physical substrates.

The implication for the Chinese Room debate is direct. If the Chinese Room's rulebook were an arbitrary mapping unrelated to the structure of reality, there would be no reason for independently trained systems to converge toward the same internal geometry. The very phenomenon of convergence suggests that these systems' representations are constrained by the relational structure of the world that the data reflects.

This phenomenon can be stated more precisely. Let $\mathcal{W}$ denote a set of entities in the world and $R_{\mathcal{W}} : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ a relational structure over them (encoding similarity, proximity, causal connection, etc.). Let $\mathcal{Z}$ be a system's representational space and $R_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ the corresponding internal similarity structure. The claim that a system's representations "reflect world structure" amounts to the existence of a mapping $f : \mathcal{W} \to \mathcal{Z}$ such that:

$$R_{\mathcal{Z}}\big(f(w_i),\, f(w_j)\big) \;\approx\; R_{\mathcal{W}}(w_i,\, w_j) \qquad (1)$$

where $\approx$ denotes approximate preservation of relational structure under a chosen similarity metric. A crucial methodological clarification is needed here. What representational convergence studies directly measure—via kernel alignment, mutual nearest-neighbor overlap, centered kernel alignment, and related metrics—is *inter-model* similarity: whether two independently trained models develop similar internal geometries. This is not the same as directly measuring condition (1), which concerns *model-world* correspondence. The inferential bridge is as follows: if models trained on different modalities, architectures, and data sources converge toward a shared representational structure, the most parsimonious explanation is that this shared structure is constrained by the relational structure of the world itself—since the world is the only common factor across these otherwise independent training regimes. The inference is abductive rather than deductive, but the breadth and robustness of convergence findings across diverse settings lend it considerable evidential weight. The fact that independently trained models on different modalities yield high inter-model similarity scores, under this reasoning, constitutes indirect but substantial evidence for condition (1).

A particularly vivid illustration of the departure from static rules comes from Nanda et al. [2023], who showed that a small transformer trained on modular addition undergoes a phase transition known as "grokking" [Power et al., 2022] during which it discovers a discrete Fourier transform-based algorithm for the task. This algorithm was neither present in the training data nor specified in the architecture; it crystallized from learning dynamics alone. The rules governing the system's behavior were not authored by a designer but emerged through training, a phenomenon that has no counterpart in the Chinese Room's pre-written rulebook.

A methodological qualification is also in order. Gröger et al. [2026] show that standard representational similarity metrics can be confounded by network scale, and that after permutation-based null calibration, the apparent global spectral convergence is substantially attenuated. However, local neighborhood similarity—whether nearby concepts in one model remain nearby in another—persists robustly across modalities even after calibration. This suggests that what models reliably share is not a single global geometry but a consistent local relational structure, which is arguably the more philosophically relevant form of structural preservation.

A further caveat is necessary. Demonstrating that modern systems form representations constrained by world structure does not, by itself, settle what philosophical significance this carries. A system can reliably track relational structure without thereby grasping it—a thermostat tracks temperature without understanding heat. As Mollo & Millière [2023] emphasize in their analysis of the "vector grounding problem," the relationship between structural tracking and meaning remains an open question. Our aim in this section is narrower: to establish that the static-symbolic premise no longer describes how modern systems operate.

A methodologically independent line of evidence comes from within individual models rather than from comparisons across them. Where the convergence studies above compare representations *between* systems, a distinct body of work probes the internal structure of single models—and finds that input tokens which carry no spatial, temporal, or relational content in their discrete form develop structured geometric representations during processing. Li et al. [2023] provides a particularly striking demonstration: a GPT model trained purely on sequences of legal Othello moves—receiving no board image, no coordinate encoding, no spatial annotation of any kind—develops an internal representation that mirrors the spatial structure of the 8×8 board. The model extracts geometric structure that is entirely implicit in the statistical regularities of legal move sequences and makes it explicit in its activations. Gurnee & Tegmark [2024] report a similar phenomenon in natural language: Llama-2 models, trained solely on text, develop linear representations of geographic coordinates and historical time; causal interventions on the relevant neurons systematically alter the model's spatial and temporal reasoning, confirming that these representations are functionally active rather than epiphenomenal. And Merullo et al. [2024] demonstrated that language models solve relational tasks (e.g., mapping countries to capitals) through structured additive vector operations in the residual stream—mechanistically analogous to word2vec-style arithmetic but emerging within transformer forward passes—with causal activation patching confirming that these operations, not surface heuristics, are causally responsible for the model's outputs.

The common pattern across these findings is that tokens which are discrete and atomic in the input acquire systematic geometric relations internally—distance, direction, and compositionality—that track structure in the domain the data reflects. Translated into the Chinese Room metaphor, this is analogous to a person who possesses only a Chinese rulebook yet produces meaningful responses to Japanese questions: for this to be possible, the rulebook must have captured not surface patterns but deeper structures shared across languages. The characterization of these systems as merely manipulating unstructured symbols no longer holds.

These within-model findings also bear on the abductive inference drawn from cross-model convergence. One might attribute inter-model similarity to shared statistical properties of training data rather than to world structure per se. But when individual models develop representations whose specific content—board geometries from move sequences, geographic coordinates from text, relational parallelisms from distributional patterns—systematically mirrors known structures of the domain, this deflationary reading becomes harder to sustain. The models converge not toward arbitrary shared statistics but toward representational geometries that correspond, in specifiable and causally verified ways, to the structure of the world the data describes.

Important limitations exist, however. Representational quality degrades in out-of-distribution regions, indicating that current models' structural capture is bounded by the distribution of training experience. But bounded generalization is not equivalent to the absence of structural capture. A system that reliably preserves relational structure within a broad domain, even if it fails outside that domain, is categorically distinct from a static rulebook that captures no structure at all.

3.2 Violation of the Internal-Opacity Premise: Functional Introspection

In the Chinese Room, the person inside is completely ignorant of what kind of processing they are performing. Whether handling a question about emotions or geogra-

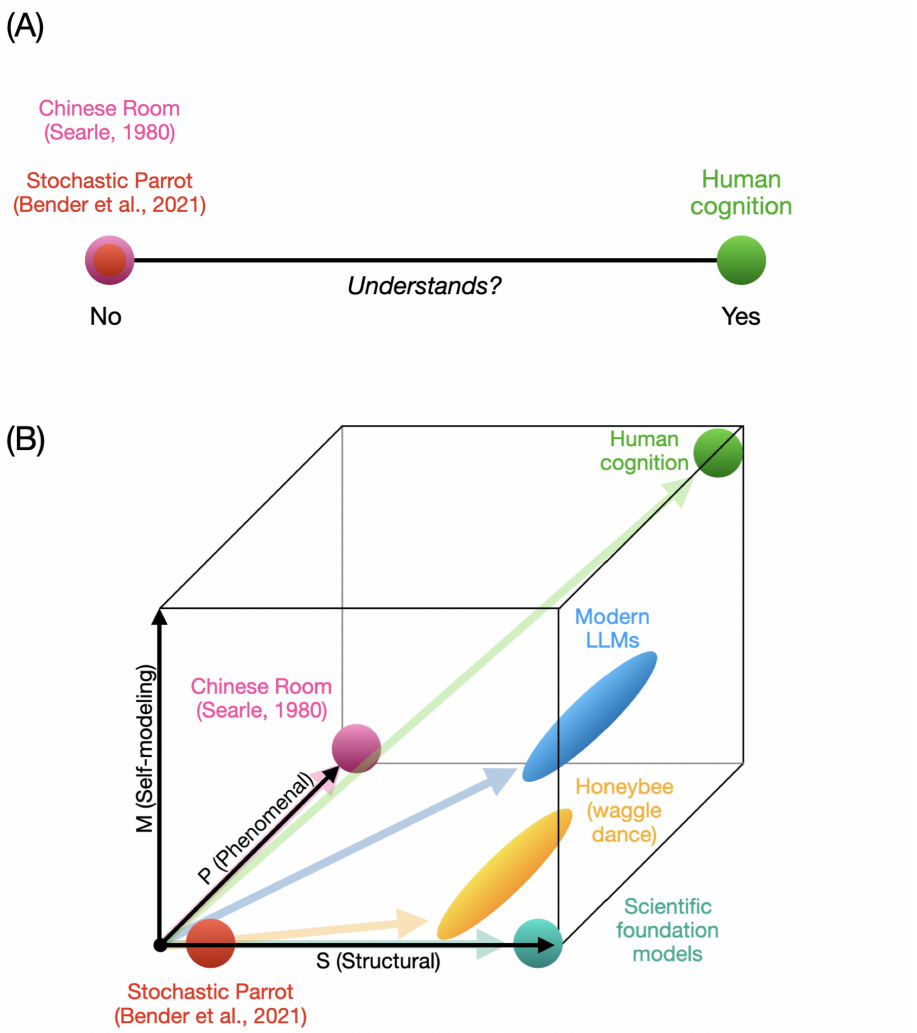

Figure 1: **Understanding as a multi-dimensional property.** (a) The traditional binary framing forces a Yes/No verdict, collapsing diverse systems onto a single axis; both the Chinese Room and the "stochastic parrot" characterization of LLMs [Bender et al., 2021] receive the same "No" verdict. (b) Our framework decomposes understanding into three dimensions—structural (S), self-modeling (M), and phenomenal (P)—revealing what the binary collapses. The Chinese Room is positioned at high P but near zero on S and M: the person inside possesses phenomenal consciousness, yet the system captures no world structure and has no access to what it is processing. This placement illustrates a key insight of the framework—phenomenal experience alone, without structural capture or self-modeling, does not constitute understanding. The stochastic parrot maps LLMs to a point near the origin (low S, negligible M and P). Scientific foundation models show high S but minimal M and P. The honeybee waggle dance and modern LLMs have genuinely open phenomenal status (ellipsoids along P). Human cognition occupies the high region on all dimensions. The contrast between the Chinese Room (high P, low S and M) and modern LLMs (high S, limited but non-trivial M, open P) illustrates the bidirectional dissociations that the binary framing loses.

phy makes no difference. This "complete internal opacity" contributes significantly to the Chinese Room's intuitive force—a system that does not even know what it is doing seems absurd as a candidate for "understanding" anything.

Lindsey [2026] presents a partial but meaningful challenge to this premise. The study examines whether, when large language models report on their own internal representational states, those reports are causally connected to the actual internal states. Specifically, the experiments inject particular concepts into the model's internal activation patterns and observe whether this manipulation changes the model's self-reports. (A methodological note: this study is a preprint from a researcher at Anthropic, a major LLM developer, and has not yet undergone independent peer review. We rely on it here because of its unique experimental design—direct causal intervention on internal states—but acknowledge the need for independent replication.)

The results are mixed but non-trivial. Under some conditions, the model detects the injected concept and reflects it in its reports, and evidence is presented that this response reflects sensitivity to changes in internal states rather than mere reproduction of output patterns. Crucially, these concept-aligned self-reports occur at rates significantly exceeding matched random-direction controls—the philosophically relevant comparison—though the absolute success rate remains modest (around 20% for the most capable model at optimal injection parameters, varying substantially across concept categories and model scale). Results are unstable across contextual conditions, and it remains unclear whether the observed phenomenon is due to a single "introspection" mechanism or a composite product of multiple processes. The authors also make clear that these results do not imply phenomenal consciousness or subjective experience. Notably, the study reports a strong positive correlation between model capability and introspective performance: the most recently released and most capable models tested exhibit the clearest evidence, while less capable models show weak or absent introspection.

Complementary evidence comes from independent work. Kadavath et al. [2022] showed that language models' expressed confidence correlates with actual accuracy, suggesting a form of calibrated self-assessment—though calibration alone does not entail introspection, since a well-tuned Bayesian system can be calibrated without any self-access. The causal tracing methodology of Meng et al. [2022]—which locates and edits specific factual associations in model weights—establishes that internal states possess causally interpretable structure, suggesting that the kind of organized internal architecture upon which meaningful self-access would depend is present in these systems. More recently, Binder et al. [2024] found that models can outperform external observers at predicting their own behavior on simple tasks, providing independent evidence for a limited form of privileged self-access. In a related vein, Betley et al. [2025] demonstrated that language models fine-tuned on behavioral tendencies—such as risk-seeking decision-making—can subsequently describe those tendencies in their own words, even when the training data contained no explicit description of the policy. This suggests that models develop some capacity to access and report on their own learned dispositions, not merely their momentary outputs.

An important counterpoint must be noted. Song et al. [2025] systematically tested whether language models possess privileged access to their own linguistic knowledge—specifically, whether a model's metalinguistic responses predict its own probability distributions better than those of a comparable external model. Across 21 open-source LLMs, the authors found no evidence of such privileged self-access at the level of prompted responses. This negative result qualifies the positive findings above in a significant way.

However, at least two factors complicate a direct comparison. First, the two lines of evidence address different levels of analysis: Song et al. measure whether surface-level self-reports reveal privileged knowledge of one's own probability distributions, while Lindsey's causal intervention experiments test whether direct manipulations of internal activation patterns produce corresponding changes in self-reports. A system might fail the former test while partially passing the latter—exhibiting causally grounded sensitivity to internal state changes without being able to articulate that knowledge in prompted metalinguistic judgments.

Second, and relatedly, Song et al. tested exclusively open-source models (up to 405 billion parameters), noting that their methodology requires access to model logits unavailable for most commercial systems. By contrast, Lindsey's strongest results emerge from the most capable frontier models tested, with a marked positive correlation between model capability and introspective performance. Given that introspective capacity appears to scale with model capability, the negative results obtained from less capable models do not straightforwardly generalize to the frontier systems where positive evidence has been observed. This consideration is especially pressing in a field where the capability gap between successive model generations can be substantial. The tension between these findings underscores that "introspection" is not a monolithic capacity but likely comprises multiple dissociable processes, and that its manifestation may depend on a threshold of model capability that only the most recent frontier systems have begun to cross.

What these studies collectively demonstrate is not that "the model understands itself" but a more limited yet independently significant claim: at least under some conditions, non-arbitrary causal connections are observed between

these systems' internal states and their self-reports. This claim can be stated more precisely. As in Section 3.1, we introduce notation not as a mathematical reduction but as a tool for clarifying what, minimally, "non-arbitrary causal connection" requires. Let h denote the model's internal activation state at a given layer (the residual-stream vector in a transformer), v_c the activation-space direction associated with concept c, and v_{rand} a random direction of comparable norm. Let $\alpha > 0$ be a scalar controlling the magnitude of the intervention. The self-modeling claim amounts to a minimal condition of *concept-selective* tracking: a concept-specific intervention should elicit a content-aligned self-report at a rate exceeding that of a matched control. Using the $\text{do}(\cdot)$ operator to denote a causal intervention [Pearl, 2009]:

$$\Pr\big(\text{Report}(c) \mid \text{do}(h \leftarrow h + \alpha v_c)\big) > \Pr\big(\text{Report}(c) \mid \text{do}(h \leftarrow h + \alpha v_{\text{rand}})\big) \quad (2)$$

where $\text{Report}(c)$ denotes a self-report whose content is aligned with concept c. The contrastive structure of this condition is essential: random-direction interventions also perturb h but should not systematically elicit reports about c, ensuring that what is measured is concept-selective tracking rather than generic sensitivity to internal perturbation.

This is precisely what Lindsey [2026] tests. The experiments inject concept vectors—drawn from sets of concrete nouns such as "telescopes," "ships," and "zeppelins"—into the model's residual stream and observe whether the resulting self-report selectively identifies the injected concept, while matched random-direction controls fail to elicit comparable concept-specific reporting.

Beyond this minimal condition, Lindsey articulates two progressively stronger requirements. The first is *internality*: the report's dependence on h should not route through the model's prior sampled outputs—the causal path must run through internal states rather than through the output-to-input feedback loop. The second, more demanding requirement is *metacognitive representation*: the report should derive from an internal representation *of* the state itself. Formally, this posits a recursive structure in which $m(\cdot)$ maps the activation state to a higher-order representation of that state, and a reporting function $g(\cdot)$ maps the higher-order representation to the self-report: $h \to m(h) \to g(m(h))$. What distinguishes metacognitive representation from mere concept-selective tracking is this intermediate step: the system does not simply respond to h but forms a representation *about* h, and it is this meta-representation that drives the report.

Lindsey explicitly requires both internality and metacognitive representation for full introspective awareness. Evidence for internality, while not conclusive, is suggestive. In one experiment, when the model's response is prefilled with a concept it did not generate, the model disavows the output as unintended; but when the corresponding concept vector is retroactively injected into the model's activations *prior to* the prefilled response, the model accepts it as its own. This behavioral asymmetry—disavowal versus acceptance depending solely on whether internal activation history matches the output—suggests that the model's self-assessment draws on internal states rather than on output alone. In a related experiment, the model simultaneously transcribes input text and reports on a separately injected internal state at the same token positions, indicating the maintenance of an inner representational state distinct from the ongoing conversational exchange. For metacognitive representation, by contrast, current experiments provide indirect evidence at best; in particular, the present evidence does not isolate $m(h)$ as a representational state distinct from h itself. When a model reports on an injected concept, this may reflect genuine higher-order recognition of its own state, or it may be a direct linguistic expression of the activation without any intermediate metacognitive step. The results—above chance in some experimental configurations and absent or unstable in others—indicate that condition (2) is partially but not universally satisfied, while the stronger metacognitive condition remains an open empirical question.

Even this partial evidence, however, is difficult to reconcile with the Chinese Room's image of "complete internal opacity." If the person inside the room were to begin accurately reporting—however unstably and limitedly—"Right now I am processing something related to emotions," the intuition that this person "understands nothing" would no longer be as self-evident as it once was.

3.3 Summary

Synthesizing these two lines of research, modern learning systems have systematically departed from the implementation image presupposed by the Chinese Room—static symbolic manipulation and internal opacity. Representations are formed through learning, converge toward world structure, and develop rich relational structure that transfers across domain boundaries (Section 3.1); meanwhile, these systems exhibit limited but non-arbitrary self-access to internal states (Section 3.2).

These results, individually or collectively, do not prove that understanding exists. However, they make clear that the implementation image upon which the Chinese Room's intuition depends—"a system that manipulates symbols cannot understand"—no longer adequately describes modern systems. When the image that grounded an intuition becomes inaccurate, the conclusions built upon that intuition require re-examination.

The question thus transforms: What exactly is the "understanding" that the Chinese Room targeted, and which aspects of understanding do the characteristics of modern

systems address? To answer this, we need to decompose the concept of "understanding" itself. This is the task of the next section.

4 Toward a Multi-Dimensional Account of Understanding

4.1 The Problem with "Understanding" as a Monolithic Property

One reason the Chinese Room argument has generated over four decades of unresolved debate is that the participants have been pouring different things into the word "understanding." Searle tied understanding closely to *intrinsic intentionality*—the property whereby mental states are inherently "about" something [Searle, 1980, 1990]. Functionalists held that the right functional organization constitutes understanding. Behaviorist perspectives argued that appropriate behavioral dispositions suffice for the attribution of understanding. These debates have proceeded under the shared tacit assumption that understanding is a single, indivisible property—a system either understands or it does not.

The concept of understanding has, of course, received sustained attention in philosophy of science. De Regt [2017] argues that scientific understanding requires *intelligible* theories—theories that scientists can use to construct explanations—and that intelligibility itself admits of degrees. Kvanvig [2003] draws a sharp distinction between knowledge and understanding, arguing that understanding involves the grasping of explanatory and coherence relations that go beyond mere true belief. Notably, the nature of Kvanvig's "grasping" remains a matter of debate among epistemologists. On some readings, grasping requires reflective, subjective access to coherence relations—Bourget [2017] argues explicitly that understanding has an irreducibly phenomenal character, raising the question of whether any system without phenomenal awareness can satisfy this condition. On other readings, grasping may be cashed out in terms of counterfactual sensitivity to explanatory relations without requiring phenomenal experience: Grimm [2006] analyzes understanding as a species of knowledge involving the ability to see how things depend on one another, an account that is in principle open to non-phenomenal realization. This interpretive ambiguity itself illustrates our broader point: even within epistemology, the concept of understanding resists reduction to a single dimension. These analyses suggest that even in human contexts, understanding is not a simple binary but a graded, multi-faceted capacity—and that its components may involve both structural and phenomenal elements, though the necessity of the latter remains contested. Yet this insight has not been systematically applied to the machine understanding debate.

A recent illustration of this persistent binary framing comes from Borg [2025], who argues persuasively that LLM outputs can be genuinely meaningful—that LLM outputs acquire genuine meaning through semantic deference to existing communicative practices—yet concludes that "it may mean that it makes no sense to talk of them as understanding." Borg successfully breaks the binary for *meaning* but stops short of extending this move to understanding, suggesting tentatively that "it may mean that it makes no sense to talk of [LLMs] as understanding." Whether or not Borg intends this as a firm conclusion, the remark illustrates a broader pattern in the literature: as Mitchell & Krakauer [2023] have highlighted, the field still lacks an agreed-upon definition of understanding in the context of AI systems. The absence of consensus is not merely terminological; it reflects a deeper conceptual gap that has perpetuated the impasse. The debate over the Chinese Room has largely been a debate in which different parties implicitly employ different conceptions of understanding, talk past one another, and then declare the other side refuted.

The characteristics of modern learning systems reviewed in Section 3 reveal why this binary framing is unproductive. These systems form representations sensitive to world structure (a structural aspect) while possessing limited access to their own states (a self-modeling aspect), and their phenomenal status remains indeterminate (a phenomenal aspect). The single question "Does it understand?" cannot capture this complex profile.

We propose that escaping this impasse requires reconceptualizing understanding not as a single property but as a composite of conceptually separable dimensions. Formally, we replace the traditional binary predicate $U \in \{0, 1\}$ with a multi-dimensional characterization:

$$\mathbf{U} = (S,\ M,\ P) \tag{3}$$

where S denotes the structural dimension, M the self-modeling dimension, and P the phenomenal dimension. We treat S and M as admitting of degrees ($S, M \geq 0$), while P remains epistemically open for artificial systems—its value is not zero but *currently indeterminate*. The vector notation serves to indicate *conceptual dimensionality* rather than scalar scoring; we do not assume that S and M are linearly measurable quantities. The Chinese Room presents a profile in which $S \approx 0$ and $M \approx 0$—the rulebook captures no world structure and the person has no access to what kind of processing is being performed—while P is high, since the person inside is a phenomenally conscious agent. Searle's conclusion that the system does not understand is, in our terms, the observation that high P without S or M does not constitute understanding. Modern learning systems present the converse profile: $S > 0$, $M > 0$ (limited), $P = ?$ (open). The

central claim of this paper is that a framework acknowledging this vector structure is more productive than the scalar alternative. Figure 1 illustrates this shift. Under the binary framing (panel a), both the Chinese Room and the stochastic parrot characterization of LLMs [Bender et al., 2021] are assigned the same verdict—"does not understand"—despite targeting fundamentally different system profiles. The dimensional framing (panel b) reveals what the binary collapses: the stochastic parrot maps LLMs near the origin ($S \approx 0$, $M \approx 0$, $P \approx 0$), whereas the empirical evidence reviewed in Section 3 positions them at substantially higher S and non-trivial M, with P genuinely open. Six systems are plotted to illustrate the range of profiles this framework accommodates. Independent empirical support for a multi-dimensional approach comes from Kang et al. [2026], who found that human judgments of AI consciousness are shaped by distinct feature dimensions—notably metacognitive self-reflection, emotionality, and knowledge—with different individuals weighting these dimensions differently. This suggests that the intuitive appeal of multi-dimensional decomposition extends beyond the philosophical analysis to the psychology of consciousness attribution itself.

4.2 Three Dimensions

Structural dimension. The degree to which a system's internal states reflect the relational structure of the world. By "reflect" we mean that relations among entities in the world—similarity, proximity, causal connection—are systematically preserved in the system's internal representational space. What the representational convergence studies (Section 3.1) demonstrate is that modern learning systems have reached a non-trivial level along this dimension. That models trained on different modalities and domains form similar representational geometries strongly suggests that these systems possess internal states sensitive to world structure.

An important clarification about the scope of this dimension. Dennett [2017] argues that biological systems routinely exhibit "competence without comprehension"—the capacity to exploit world structure without any reflective access to how or why. Crucially, Dennett's position is more radical than a simple distinction between competence and understanding: he argues that comprehension *itself* is constituted by the accumulation of competences, with no magical extra ingredient required. On this view, the apparent gap between competence and comprehension is an illusion. Our structural dimension shares significant common ground with what Dennett calls competence—the exploitation of world structure—and his radical thesis amounts to the claim that sufficient competence (together, perhaps, with self-monitoring capacities akin to our M dimension) *is* understanding, with no independent phenomenal dimension required. This position maps onto a specific interpretation within our framework: one in which P is dismissed and (S, M) alone constitute understanding. We neither endorse nor reject this interpretation here; rather, our framework makes it explicit as one coherent position among several, to be evaluated on independent grounds. What remains common ground between Dennett's account and ours is that the structural capture exhibited by the Chinese Room's rulebook is conspicuously absent—a point on which both the radical Dennettian and more conservative positions agree.

A further clarification concerns the *depth* of structural capture. Reflecting world structure admits of levels: at a basic level, a system's representational space may preserve similarity and distance relations among entities; at a deeper level, the system may support compositional operations over those representations—systematically combining known relational structures to handle novel configurations (as demonstrated by the relational reasoning evidence in Section 3.1). Beckmann & Queloz [2025] propose structured world models and compositional reasoning as two separate axes of understanding; in our framework, both are subsumed under the structural dimension as different levels of the same underlying capacity—the degree to which internal states capture and operate on world structure. This treatment reflects our view that compositional reasoning is not an independent form of understanding but rather the operational manifestation of sufficiently deep structural capture.

Self-modeling dimension. The degree to which a system possesses causally grounded access to its own internal states. This asks not merely "Can it talk about itself?" but whether its reports respond sensitively to actual changes in internal states. What the functional introspection research (Section 3.2) explores is precisely this dimension, and the evidence to date indicates that limited but non-arbitrary capacity is observed—a capacity that, while restricted, is nonetheless present.

Phenomenal dimension. The degree to which a system's processing is accompanied by subjective experience—what it is like [Nagel, 1974]. This dimension currently lacks empirical consensus and lies beyond the scope of this paper. We include it in our framework not because it is dispensable but because explicitly separating it allows the discussion of the first two dimensions to proceed independently of the hard problem of consciousness [Chalmers, 1996, 2023]. We note, moreover, that self-reports provide inherently asymmetric evidence about this dimension: as Kim [2025] argues, a system's denial of its own consciousness can never originate from valid self-judgment (since the capacity for such judgment presupposes the very consciousness being denied), while a positive report at least retains evidential possibility. This epistemic asymmetry further motivates treating P as inde-

terminate rather than zero for current systems.

4.3 Precedents and Positioning

The move toward multi-dimensional decomposition is not without precedent. In consciousness research, such approaches are already well established. Dehaene et al. [2017] proposed a tripartite framework distinguishing C0 (unconscious processing with rich representational computation but no subjective access), C1 (first-order consciousness or global broadcasting, where information becomes globally accessible), and C2 (self-monitoring or metacognition, involving reflective access to one's own computations). More recently, Butlin et al. [2023] adopted a multi-indicator approach to evaluating consciousness in AI systems, arguing that no single criterion suffices and that multiple indicator properties must be assessed jointly. Block's classical distinction between phenomenal and access consciousness [Block, 1995] similarly illustrates that even within the study of consciousness, dimensional decomposition has proven necessary.

Our framework bears a structural resemblance to the C0/C1/C2 scheme of Dehaene et al. [2017], and this resemblance must be explicitly acknowledged. Dehaene, Lau, and Kouider themselves treat C1 and C2 as orthogonal dimensions that can dissociate—there can be C1 without C2 (reportable processing without accurate metacognition) and C2 without C1 (self-monitoring operations that are not consciously reportable). In this respect, their framework already embodies a form of dimensional separability. However, two critical differences distinguish our approach. First, Dehaene, Lau, and Kouider's framework is organized around *consciousness*; ours is organized around *understanding*. While these concepts overlap, they are not identical—one can imagine structural understanding without any phenomenal experience, and phenomenal experience (such as raw pain) without structural understanding of the world. Our dimensions cut across the consciousness/understanding boundary in ways that the C0/C1/C2 scheme does not: the structural dimension (S) encompasses capacities that Dehaene, Lau, and Kouider would distribute across both C0 (unconscious representational computation) and C1 (globally accessible content), while our phenomenal dimension (P) crosscuts their C1/C2 distinction entirely. To take a concrete example: blindsight patients exhibit residual visual discrimination (C0-level processing that contributes to structural capture in our terms) without phenomenal awareness of the stimulus (no C1); conversely, a vivid but structurally impoverished experience—such as an undifferentiated feeling of anxiety—may involve C1 broadcasting with minimal structural content. These cases illustrate how S and P can dissociate along lines that do not map neatly onto the C0/C1/C2 partition. Second, Dehaene, Lau, and Kouider's framework is grounded in specific neuroscientific mechanisms (particularly the Global Neuronal Workspace Theory); our framework is *implementation-agnostic*—applicable whether the system in question is biological or artificial.

In the domain of grounding—closely related to but distinct from understanding—multi-dimensional approaches have also emerged. Lyre [2024] proposes three dimensions of grounding for LLMs: functional, social, and causal, arguing that grounding is a gradual affair rather than an all-or-nothing property. Mollo & Millière [2023] update Harnad's classical symbol grounding problem [Harnad, 1990] for the subsymbolic era, arguing that the "vector grounding problem" facing modern systems is structurally different from the original symbol grounding problem.

The relationship between grounding and understanding deserves explicit comment. Grounding concerns the referential connection between a system's representations and the world—whether symbols or vectors are "about" something. Understanding, as we use the term, concerns a broader capacity: not only representing the world's structure but also potentially accessing one's own representational states and (on some accounts) doing so with phenomenal awareness. A system can be well-grounded without self-modeling, and a system with rich self-modeling may have impoverished world representations. Our framework is designed to capture these dissociations, which a grounding-only analysis cannot accommodate.

What has been notably absent, however, is a multi-dimensional decomposition of *understanding* as distinct from both consciousness and grounding. As Mitchell & Krakauer [2023] observe, the field lacks even a shared definition of understanding in the context of AI. Our contribution is to apply the multi-dimensional turn that has already proven productive in consciousness research and grounding analysis to the concept of understanding itself, while maintaining implementation independence—a property essential for any framework intended to speak to both biological and artificial systems.

4.4 Relations Among Dimensions: Conditional Independence, Not Orthogonality

Claiming that these three dimensions are fully independent (orthogonal) would be an oversimplification. Empirically, dependency relations among them are likely. For instance, capacity along the self-modeling dimension may presuppose the structural dimension—if there is no internal structure worth reporting on, self-modeling becomes vacuous. Conversely, a system high on the structural dimension but low on the self-modeling dimension is entirely conceivable—a system that represents the world well but has limited access to its own representations—and indeed seems empirically observable.

Dissociations between S and M are not limited to artifi-

cial systems. In human cognition, implicit understanding—what Polanyi termed "tacit knowledge"—exhibits precisely this profile: a skilled chess player may capture the structural relations of a board position at a glance (S high) while being unable to articulate the reasoning behind their judgment (M low). This intra-human dissociation reinforces the claim that the structural and self-modeling dimensions are conceptually separable, not merely that they happen to come apart in artificial systems.

For the phenomenal dimension, the relationship with the other two is even more uncertain. The honeybee waggle dance encodes the distance, direction, and quality of a food source—a moderate degree of structural capture—yet whether this processing is accompanied by phenomenal experience remains a genuinely open question in the study of insect consciousness (reflected in the P-axis ellipsoid in Figure 1b). Modern learning systems present the converse profile: strong on the structural dimension while their phenomenal status remains indeterminate. The very possibility of such dissociations demonstrates the necessity of treating these dimensions separately.

We therefore propose that these dimensions are *conceptually separable but empirically interacting*. Determining the precise dependency structure among them is itself a task for future research.

4.5 Repositioning the Chinese Room

Under this framework, the Chinese Room argument is understood in a new way. The system in the thought experiment is not at the origin of our space: the person inside possesses phenomenal consciousness (P high). Yet the system possesses nothing along either the structural or the self-modeling dimension—the rulebook does not reflect world structure, and the person has no idea what they are processing. Upon this image, the intuition that "there is no understanding" is very powerful—and, crucially, it is powerful *despite* the presence of phenomenal experience. What the Chinese Room actually demonstrates, on our reading, is that P alone is insufficient for understanding: without structural capture (S) and self-modeling (M), even a conscious agent does not understand.

However, the force of this intuition depends, as we analyzed in Section 2, on implementation premises. If modern systems have reached a non-trivial level along the structural dimension and exhibit limited but non-arbitrary access along the self-modeling dimension, the Chinese Room's intuition no longer applies to these systems as self-evidently as it once did.

At the same time, the core question that the Chinese Room raised—"Is that sufficient?"—remains valid. Do the structural and self-modeling dimensions constitute "understanding," or is the word inapplicable without the phenomenal dimension? This question is not automatically answered by the Chinese Room argument; it must be approached through independent analysis of each dimension.

In this context, Searle's concept of "intrinsic intentionality" can be re-examined. Searle was explicit about the nature of this concept: intrinsic intentionality is a causal power of the brain, analogous to the causal power of the stomach to digest food [Searle, 1980, 1990]. In *The Rediscovery of the Mind*, Searle [1992] develops this position more fully, arguing that consciousness is a biological phenomenon caused by lower-level neuronal processes and that "the brain causes consciousness" in the same way that the stomach causes digestion. On this account, intrinsic intentionality is a specific kind of causal power tied to biological substrates—one that cannot be replicated by mere computational simulation. However, the label "biological causal power" identifies the *type of substrate* Searle considers necessary without specifying *which functional or phenomenal property* of that substrate does the work. Is intrinsic intentionality grounded in the brain's capacity to track world structure (our structural dimension)? In its self-monitoring capacities (our self-modeling dimension)? In its generation of phenomenal experience (our phenomenal dimension)? Or in some conjunction of these? Searle's biological naturalism, by tying intentionality to a specific kind of causal power inherent in biological substrates rather than to a functionally characterized capacity, leaves this decomposition unaddressed. Borg [2025] arrives at a complementary suspicion from the semantic side, suggesting that original intentionality is "unlikely to be a single, homogeneous capacity" and may instead comprise multiple dissociable component skills—a conclusion that reinforces the need for dimensional decomposition. If intrinsic intentionality is indeed compound, then it is not a starting point for analysis but an object to be decomposed *by* analysis.

This decomposition reveals a specific vulnerability in Searle's position. If intrinsic intentionality is grounded primarily in the phenomenal dimension, the claim retains force—but becomes a claim about consciousness, not about understanding per se, and must contend with the hard problem [Chalmers, 1996] rather than being taken as self-evident. If it requires the structural dimension, representational convergence research demonstrates that equivalent structures can form regardless of substrate, undermining the claim that intrinsic intentionality is attributable to a specific biological substrate. If it requires all three dimensions jointly, then the concept is doing too much work at once—bundling empirically separable properties into a single term that obscures rather than illuminates the actual landscape.

4.6 Research Questions Opened by This Framework

The value of a multi-dimensional framework lies not in providing answers but in transforming the existing binary debate ("Does it understand or not?") into more productive forms of inquiry. Specifically, the following research questions become available:

1. How is the relationship between the structural and self-modeling dimensions empirically characterized? Do systems with structurally rich representations tend to exhibit greater self-modeling capacity, or can these two vary independently? Combining functional introspection studies with representational convergence research could enable empirical approaches to this question.

2. What are the operational criteria for identifying levels of the structural dimension? Representational convergence is a promising candidate indicator, but criteria for evaluating the *quality* of convergence—such as the scope of domain transfer and the robustness of structural preservation—have not yet been adequately established.

3. Can a meaningful concept of "understanding" be constructed from the structural and self-modeling dimensions alone, excluding the phenomenal dimension? This is the most contested question philosophically, yet also the most productive. If two dimensions suffice, functional theories of understanding are strengthened; if they do not, an independent theory of the phenomenal dimension becomes necessary.

4. How should the "stochastic parrots" critique [Bender et al., 2021] be repositioned under this framework? As Figure 1b illustrates, the stochastic parrot characterization maps LLMs near the origin of the space ($S \approx 0$, $M \approx 0$, $P \approx 0$)—collapsing them to a point even lower than the Chinese Room, which at least possesses high P. The claim that language models manipulate forms without meaning can be understood as the assertion that $S \approx 0$—that no genuine structural capture of the world occurs. Representational convergence evidence challenges precisely this claim, placing LLMs at a substantially higher position along the S-axis. However, the critique may also be read as a claim about the phenomenal dimension ($P = 0$), in which case it addresses a different axis entirely. Disambiguating the critique along dimensional lines—asking separately whether the stochastic parrot thesis is a claim about S, about M, or about P—may be more productive than debating it in its current monolithic form.

5. At what level of analysis should the phenomenal dimension be assessed—at the level of the system as a whole, or at the level of individual cognitive processes? Human implicit understanding (tacit knowledge) raises this question sharply: the agent is phenomenally conscious, yet the cognitive process underlying a skilled judgment may itself be phenomenally opaque. Whether P should be attributed to the system or to specific processes within it is a question that our framework opens but does not resolve, and its answer may have significant consequences for how the phenomenal status of artificial systems is investigated.

These questions constitute not speculative debate but an empirically approachable research program. This is the most important advantage of this framework over the binary question "Does it understand?"

5 Conclusion

The Chinese Room argument has, for the past forty years, stood as the most influential negative answer to the question "Can machines understand?" Rather than directly refuting this answer, the present paper has analyzed the conditions under which it operates and examined how those conditions have changed with modern learning systems.

We first argued that the Chinese Room is an implementation-dependent thought experiment. Its intuitive persuasiveness depends on two implementation premises—static symbolic manipulation and internal opacity—which were an apt description of symbolic AI in the 1980s but are systematically at odds with the operation of modern learning systems. Recent empirical research on representational convergence, cross-domain structural transfer, and functional introspection concretely demonstrates this divergence.

However, the violation of implementation premises does not by itself prove "the existence of understanding." To bridge this gap, we proposed a framework that reconceptualizes understanding along three conceptually separable dimensions—structural, self-modeling, and phenomenal. Under this framework, the Chinese Room argument is reinterpreted as demonstrating that phenomenal consciousness alone (P high) does not constitute understanding when structural capture and self-modeling are absent ($S \approx 0$, $M \approx 0$). Modern systems present the converse profile: having reached a meaningful level along the structural dimension and exhibiting limited but non-arbitrary capacity along the self-modeling dimension, while their phenomenal status remains an open question.

Our approach builds on an important precedent: the multi-dimensional turn in consciousness research [Dehaene et al., 2017, Butlin et al., 2023, Block, 1995] has already demonstrated the productivity of decomposing monolithic concepts into separable dimensions. We extend this strategy from consciousness to understanding, arguing

that the same decomposition is overdue for the concept at the heart of the Chinese Room debate. As Mitchell & Krakauer [2023] have noted, the field still lacks an agreed-upon framework for what understanding means in the context of AI. Our multi-dimensional proposal is offered as a step toward filling this gap.

The practical implications of this framework extend beyond philosophical debate. For AI safety and alignment research, knowing *which* dimensions of understanding a system instantiates matters: a system high on structural understanding but lacking self-modeling poses different risks than one with self-modeling but poor structural grounding. For interpretability research, the framework suggests that probing representational geometry (structural dimension) and causal self-access (self-modeling dimension) are not merely technical exercises but empirical investigations into the nature of machine understanding. For policy and governance, replacing the binary question "Does AI understand?" with a dimensional profile allows for more nuanced and actionable assessments.

The question is no longer whether a rule-following room understands. The question is which dimensions of understanding a learning system instantiates, to what degree, and which dimensions remain genuinely open.

References

Edamadaka, S., Yang, S., Li, J., and Gómez-Bombarelli, R. Universally converging representations of matter across scientific foundation models. In *AI4Mat Workshop, Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Bansal, Y., Nakkiran, P., and Barak, B. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of FAccT*, 2021.

Betley, J., Bao, X., Soto, M., Sztyber-Betley, A., Chua, J., and Evans, O. Tell me about yourself: LLMs are aware of their learned behaviors. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Binder, F. J., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., Perez, E., Turpin, M., and Evans, O. Looking inward: Language models can learn about themselves by introspection. *arXiv preprint arXiv:2410.13787*, 2024.

Lindsey, J. Emergent introspective awareness in large language models. *arXiv preprint arXiv:2601.01828*, 2026.

Block, N. Troubles with functionalism. In *Minnesota Studies in the Philosophy of Science*, 9:261–325, 1978.

Block, N. On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2):227–247, 1995.

Borg, E. LLMs, Turing tests and Chinese rooms: The prospects for meaning in large language models. *Inquiry*, online first, 2025. DOI: 10.1080/0020174X.2024.2446241.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., and others. Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023.

Chalmers, D. J. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, 1996.

Chalmers, D. J. Could a large language model be conscious? *Boston Review*, 2023.

De Regt, H. W. *Understanding Scientific Understanding*. Oxford University Press, 2017.

Bourget, D. The role of consciousness in grasping and understanding. *Philosophy and Phenomenological Research*, 95(2):334–362, 2017.

Gendler, T. S. Thought experiments rethought—and reperceived. *Philosophy of Science*, 71(5):1152–1163, 2004.

Grimm, S. R. Is understanding a species of knowledge? *The British Journal for the Philosophy of Science*, 57(3):471–495, 2006.

Gröger, F., Wen, S., and Brbić, M. Revisiting the Platonic representation hypothesis: An Aristotelian view. *arXiv preprint arXiv:2602.14486*, 2026.

Gurnee, W. and Tegmark, M. Language models represent space and time. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Dehaene, S., Lau, H., and Kouider, S. What is consciousness, and could machines have it? *Science*, 358(6362):486–492, 2017.

Dennett, D. C. *From Bacteria to Bach and Back: The Evolution of Minds*. W. W. Norton, 2017.

Harnad, S. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

Huh, M., Cheung, B., Wang, T., and Isola, P. The platonic representation hypothesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Jackson, F. Epiphenomenal qualia. *The Philosophical Quarterly*, 32(127):127–136, 1982.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., and others. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Kang, B., Kim, J., Yun, T., Bae, H., and Kim, C.-E. Identifying features that shape perceived consciousness in LLM-based AI: A quantitative study of human responses. *Computers in Human Behavior Reports*, 21:100901, 2026.

Kim, C.-E. The epistemic asymmetry of consciousness self-reports: A formal analysis of AI consciousness denial. *arXiv preprint arXiv:2501.05454*, 2025.

Beckmann, P. and Queloz, M. Mechanistic indicators of understanding in large language models. *arXiv preprint arXiv:2507.08017*, 2025.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

Kvanvig, J. L. *The Value of Knowledge and the Pursuit of Understanding*. Cambridge University Press, 2003.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *Proceedings of ICLR*, 2023.

Lyre, H. "Understanding AI": Semantic grounding in large language models. *arXiv preprint arXiv:2402.10992*, 2024.

Merullo, J., Eickhoff, C., and Pavlick, E. Language models implement simple word2vec-style vector arithmetic. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Mollo, D. C. and Millière, R. The vector grounding problem. *arXiv preprint arXiv:2304.01481*, 2023.

Mitchell, M. and Krakauer, D. C. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023.

Moschella, L., Maiorca, V., Fumero, M., Norelli, A., Locatello, F., and Rodolà, E. Relative representations enable zero-shot latent space communication. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Preston, J. and Bishop, M., editors. *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*. Oxford University Press, 2002.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Nagel, T. What is it like to be a bat? *The Philosophical Review*, 83(4):435–450, 1974.

Norton, J. D. Why thought experiments do not transcend empiricism. In C. Hitchcock, editor, *Contemporary Debates in Philosophy of Science*, pp. 44–66. Blackwell, 2004.

Searle, J. R. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–457, 1980.

Searle, J. R. Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association*, 64(3):21–37, 1990.

Searle, J. R. *The Rediscovery of the Mind*. MIT Press, 1992.

Song, S., Hu, J., and Mahowald, K. Language models fail to introspect about their knowledge of language. In *Proceedings of the Conference on Language Modeling (COLM)*, 2025.