Trends in Cognitive Sciences

CellPress OPEN ACCESS

Feature review

Machine understanding

Huili Chen [1,2,*], Stephen R. Grimm [3], Olga Russakovsky [4], and Tania Lombrozo [5]

What do artificial intelligence (AI) systems "understand"? This question arises not only in assessing a system's intelligence but also in evaluation practices to ensure the safe and responsible deployment of AI. Drawing on scholarship from philosophy and cognitive science, and informed by current practices in AI, we develop a framework for asking more precise questions and making more precise claims about machine understanding. We conceptualize understanding as a relation between a system (S) and a target of understanding (T), and we discuss how to specify the relation, the system, and the target, offering a landscape of options in each case. Our goal is not to defend a particular account of understanding, but to provide conceptual tools for those working to assess or advance machine understanding.

Highlights

New developments in artificial intelligence seem to be expanding the scope of what machines 'understand': image recognition systems appear to possess some understanding of objects and scenes, while large language models appear to possess some understanding of language.

Evidence of advances (or limitations) in machine understanding is used to make claims about the safety and intelligence of artificial intelligence systems.

However, such claims require an account of 'machine understanding' that clearly specifies what constitutes understanding and how it can be evaluated.

While the fields of artificial intelligence and machine learning have not converged on an account of machine understanding, different assumptions are reflected in contemporary practice and relate to accounts of understanding from philosophy and the cognitive sciences.

Charting the landscape of machine understanding

Do large language models (LLMs) understand language? Do educational artificial intelligence (AI) systems understand the material they teach? Do therapeutic AI agents understand their patients? As AI advances, fundamental questions emerge: What do AI systems understand, and how can we know? To answer these questions, we need to grapple with the term itself—what it means to "understand".

Philosophers often start with the commonsense idea that to understand is "to see how things hang together" [1] or "to see how things are connected" [2]. It seems to follow that understanding is, in some sense, holistic [3], and that understanding comes in degrees: the better one understands some target, the better one appreciates how its various elements are connected or related [4,5]. However, beyond these commonsense starting points, existing accounts of understanding rapidly diverge, and accounts of 'machine' understanding are in their infancy (see Box 1). Yet claims about what AI systems do or do not understand are pervasive in computer science [6,7] and beyond [8–12].

Achieving greater clarity and consensus concerning machine understanding is important for both practical and theoretical reasons. Despite their impressive achievements, AI systems regularly fail in ways that suggest a critical gap between surface-level pattern matching and deeper understanding [7]. Often, limitations are discovered only after consequential real-world failures, such as errors in healthcare [13] or legal decision-making[i]. Greater precision about what constitutes understanding, and how to assess it, is therefore critical for improving evaluation practices and supporting the safe and responsible deployment of AI systems—not just in healthcare [14] and law [15], but also in education [16], scientific research [11], and more.

Here, we offer a framework for asking and answering questions about machine understanding. Our framework organizes and synthesizes existing ideas across psychology, philosophy, and computer science—some of which have existed for decades and others that reflect cutting-edge developments. To do so, we introduce two families of proposals for what constitutes understanding (model based and ability based) that can be augmented with additional requirements

[1]Program in Cognitive Science, Princeton University, Princeton, NJ, USA
[2]Faculty of Information, University of Toronto, Toronto, ON, Canada
[3]Department of Philosophy, Fordham University, New York, NY, USA
[4]Department of Computer Science, Princeton University, Princeton, NJ, USA
[5]Department of Psychology, Princeton University, Princeton, NJ, USA

*Correspondence.
huili.chen@utoronto.ca (H. Chen) and lombrozo@princeton.edu (T. Lombrozo).

© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Box 1. The philosophy of understanding

Accounts of understanding have been developed in epistemology and the philosophy of science, often with the aim of capturing a particular kind of understanding (such as scientific [128] or interpersonal [129,130] understanding). Many accounts build on the premise that understanding involves seeing how things are connected [131], where the objects of understanding are structures or systems of some kind [132,133], with parts or elements that depend upon or otherwise relate to one another. However, accounts vary in how they articulate these parts and relations (the "things" and their connections). Accounts also differ in how they interpret the idea of "seeing" (or "grasping"). According to James Woodward, for example, these metaphors are ultimately grounded in the ability to answer "What if things had been different?" questions—thus being able to anticipate how changes in one element of the system will lead, or fail to lead, to changes in other elements ([134], cf. [45]). "Seeing" thus seems to have a distinctively modal profile: someone who understands a system does not just take in how the system (actually) is, but also the various possibilities the system affords, including which relationships are contingent, necessary, and so on [135].

Philosophers disagree on whether understanding requires consciousness [136]. Earlier discussions tended to dismiss the idea [137], but more recent scholars have suggested that something like the ability to imagine (i.e., the conscious experience of imagining) seems essential [138] see also [139], [140]. For example, one might understand, in some way, that the Sun is 1,300,000 times the size of the Earth, but when one is shown a scale image of the Sun versus the Earth (say, a bowling ball vs. a peppercorn), this allows for a much deeper kind of understanding. Alternatively, it might allow for an entirely different kind of understanding—perhaps an understanding that eludes nonconscious beings.

Another ongoing debate concerns whether understanding is "factive"—that is, whether the representations that make up someone's understanding need to be true or accurate [140]. "Non-factivists" note that idealizations in science (e.g., a frictionless plane, ideally rational agents) often help us understand a target system, even though they are not, strictly speaking, true or accurate [141]. "Factivists" have replied that when idealizations help understanding, it is not because of the inaccurate or idealized information [2,143]. Approaches to machine understanding can learn from these debates, and considering artificial systems can, in turn, bring new philosophical issues to light [143–146].

(etiology based and phenomenology based). These proposals provide a set of distinctions and a corresponding vocabulary to integrate prior work and guide new research moving forward.

Our framework is consistent with a pluralist approach to machine understanding—one that recognizes multiple senses of understanding that can potentially diverge. Our goal is not to endorse a particular account of understanding but to provide a set of tools for asking and answering questions about machine understanding within a common framework that enables different scholars to effectively communicate and make progress in defining, assessing, and advancing machine understanding.

Asking precise questions, making precise claims

Assessing machine understanding requires answering questions of the form, "Does S understand T?", where S represents a system and T represents a target of understanding. For instance, in asking, "Do LLMs understand language?", LLMs are the relevant system (S), which could refer to the general class of models, a trained system (such as GPT-3.5-turbo-1106), a trained system with access to additional resources (such as the internet), or even a distributed system that includes a human providing prompts (see Box 2). T could refer to word meanings, English syntax, or other aspects of language. Specifying S and T is important not only for posing precise questions but also because the relevant notion of understanding could depend on S and T. For instance, understanding a programming language could involve a different notion of understanding from understanding a person.

Typically, AI understanding is operationalized by measuring the performance of a trained system on a target benchmark. However, it is often unclear whether performance is taken to *constitute* understanding or to offer *evidence* of understanding. To illustrate why this distinction matters, consider two systems that output a prediction about the distance traveled by a falling object in the number of seconds specified by the input. The first system, LOOK-UP, implements a look-

CellPress
OPEN ACCESS

Box 2. Taxonomy of understanding systems

The landscape of AI is increasingly heterogeneous, with systems varying in their architectures, capabilities, and applications—from next-token prediction neural networks to multimodal systems [147], generative agents that simulate social behaviors [148], and LLM-powered robots executing physical tasks [149]. This heterogeneity requires a framework for evaluating machine understanding at various levels of abstraction and specification. Here, we offer distinctions between four levels to help guide claims about understanding, but this is not an exhaustive taxonomy (see Figure I).

- A 'base system' refers to core machine learning techniques and algorithms, such as neural networks, reinforcement learning, or autoregressive models [150]. An example of a question at the base systems level is, “Do LLMs understand language?”, where the system (LLMs) refers to a class of systems using transformer-based architectures, rather than a particular instantiation (such as GPT-5-202508-07). Claims about the universal properties of LLMs, including their understanding, are not uncommon. For example, some caution against claims that “large neural language models...*understand* language” (74, emphasis added).
- A 'trained system' involves one or more base systems that are instantiated and trained, but “frozen”. An example of a question at the trained systems level is, “Does ChatGPT/GPT-4 understand English?”, where ChatGPT/GPT-4 combines multiple base systems (a transformer-based LLM plus reinforcement learning from human feedback), and where the integrated system has been instantiated and trained. Claims about understanding at the trained systems level are pervasive, with benchmark leaderboards purporting to compare, for instance, trained systems’ “ability to understand and reason about texts” (e.g., SuperGLUE [42]).
- An 'embedded system' is a trained system with access to one or more dynamic resources, such as external tools [84], knowledge bases (e.g., the internet), or human instructions (e.g., chain-of-thought prompting) [151]. At this level, we can ask, “Does ChatGPT/GPT-4 understand logic problems (with a human providing chain-of-thought prompting)?” An example comes from a paper [152] using multi-step chain-of-thought prompting with a multimodal LLM “to enhance the model’s understanding of geometry problems.”
- A 'distributed system' encompasses not one trained system, but a trained system and a dynamic resource or multiple trained systems and/or resources. At this level, the unit of analysis is the distributed system, and understanding need not reside in any individual system or resource (see [153] for related points about intelligence). At this level, we can modify the previous question: “Do ChatGPT/GPT-4 and the human (who is providing chain-of-thought prompting), as a unit, understand logic problems?”

Note that many contemporary AI systems, presented as single ‘trained systems’, can, in practice, function as meta-systems that route queries across different underlying trained models, reasoning modes, or tool-use configurations. This can blur the boundary between the trained-system and distributed-system categories in our taxonomy. When architectures are opaque, the appropriate locus of understanding might have to be identified at the level of analysis accessible to researchers.

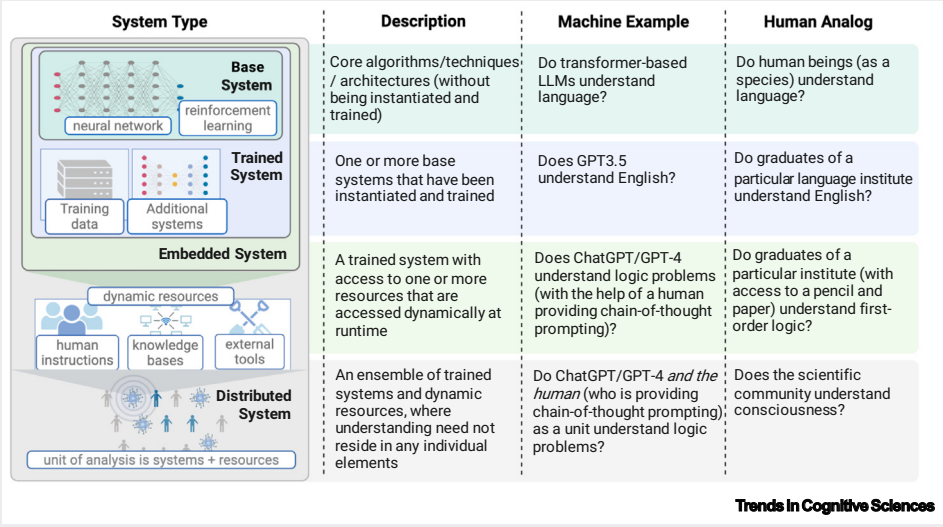

Figure I. Schematic representation of four levels of system abstraction and specification for machine understanding. GPT: generative pre-trained transformer; LLMs: large language models.

up table that includes a finite set of observations for how far an object fell in different amounts of time (e.g., 1 second, 2 seconds, etc.). On a particular assessment, it returns the correct prediction for how far an object will fall on 83% of the provided inputs. The second system, EQUATION, instead derives distance from Newtonian laws by computing a formula (d=0.5 × g × t^2). However, because the formula corresponds to an idealized version of the system (that does not, for instance, incorporate air friction), it can be less accurate than LOOK-UP on values that correspond to LOOK-UP's entries; on the same assessment, it only returns the correct answer on 76% of the inputs. Two scientists might agree on these facts yet disagree on which system has a deeper understanding of falling objects: Dr Benchmark might argue that benchmark performance partially or fully *constitutes* understanding and so LOOK-UP possesses deeper understanding than EQUATION, whereas Dr Decode could argue that benchmark performance only offers indirect *evidence* of understanding and that EQUATION possesses deeper understanding than LOOK-UP because of its internal structure: it captures something about the data-generating process that is at best implicit in the look-up table. Which scientist is correct?

The answer depends on one's account of understanding. An account of machine understanding can resolve such disputes by differentiating what constitutes understanding from what merely offers (indirect) evidence. Accounts of machine understanding can also identify what (if anything) is deficient about LOOK-UP, the system with a look-up table. In psychology and education, understanding is often differentiated from rote memorization (e.g., Bloom's Taxonomy [17]), and in computer science, understanding is sometimes contrasted with "parroting" [18] or "mere retrieval". The idea that understanding requires more than mechanical processing and retrieval is immortalized in the Chinese Room thought experiment [19], in which the philosopher John Searle invites us to imagine a person locked in a room, without any understanding of Chinese, who receives messages in Chinese and follows the steps in a book of instructions to determine appropriate Chinese responses. Searle claims that it will seem like the person understands Chinese but does not. The philosopher Ned Block offers an example that better matches contemporary debates about "mere retrieval", in which he suggests that a system that relies on a series of look-up tables might mimic intelligence but actually lack understanding [20].

Although there is disagreement about what these thought experiments show, we take it as a guiding assumption that an account of understanding should offer some insight concerning what (if anything) is deficient about a look-up table or "mere retrieval" when it comes to understanding. We, thus revisit the contrast between LOOK-UP and EQUATION for each account of understanding.

Accounts of understanding

Our framework involves two core accounts of understanding: model based and ability based. These core accounts can be combined to yield hybrid accounts, and they can incorporate additional requirements: "etiology-based" requirements that involve a system's history, and "phenomenology-based" requirements that involve subjective experience. We summarize the framework in Figure 1 and in Figure 2 we offer a five-step guide for applying the framework to evaluate claims about machine understanding.

Model-based accounts

The intuition behind model-based accounts is that understanding is a matter of what a system has "inside": particular representations, algorithms, or other internal properties. This is what motivates Dr Decode, and it reflects two common practices in AI: engineering structured representations of the target of understanding or testing for their presence. For instance, decoding techniques are often taken to reveal "world models" [21]: structure-preserving (causal)

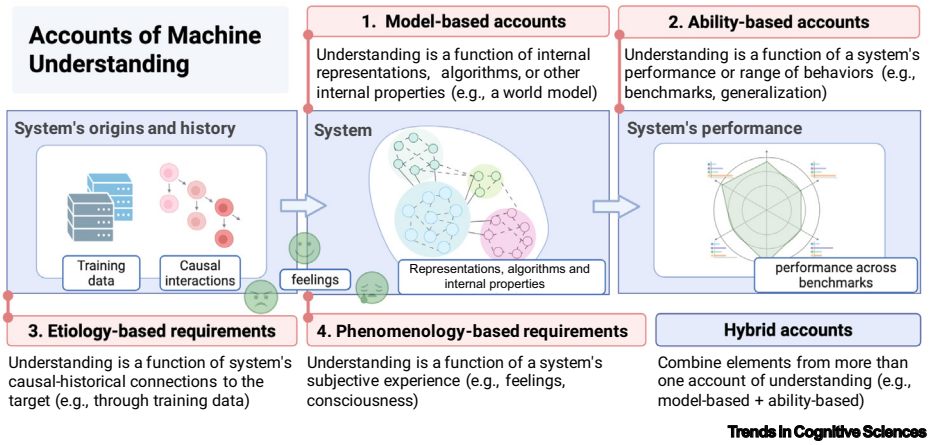

Figure 1. Framework for defining machine understanding: core accounts, additional requirements, and hybrid views. The figure presents a landscape of different notions of machine understanding. The accounts of understanding are organized into two primary categories (model based and ability based) and two sets of additional candidate requirements (etiology based and phenomenology based). On model-based accounts, understanding depends on a system's internal representations, algorithms, or other internal properties, such as a world model. On ability-based accounts, understanding depends on successful performance or patterns of behavior, such as benchmark success. Etiology-based requirements ask whether the relevant model or ability was acquired through the correct causal history or connection to the target of understanding. Phenomenology-based requirements ask whether understanding also necessitates relevant subjective experience. Etiology and phenomenology are add-on requirements that can constrain either core account (model based or ability based). Hybrid views combine elements of more than one account or requirement. This menu of options can serve as a basis for selecting a single account of understanding or for defining a set of possibilities within a pluralist approach.

representations of some target, whether it be the structure of a context-free grammar or aspects of the natural [22,23] or social [24,25] world (see Box 3). Complementing these decoding strategies, other research focuses on explicitly engineering knowledge structures, as seen in ConceptNet [26] or neuro-symbolic approaches such as knowledge-enhanced graph neural networks [27]. Across these cases, taking some internal representation or world model to constitute understanding is the core of a model-based approach.

Within philosophy, a variety of model-based accounts have been proposed for human understanding. Many accounts treat understanding as a kind of knowledge (e.g., of explanations or dependence relations [28]). Extending these accounts to machines is not straightforward, since philosophers standardly take knowledge to involve (true) beliefs [29], and it's controversial whether current AI systems possess beliefs [30–32]. Instead, such accounts could be extended to encompass (nonbelief) representations involving similar causal or explanatory content. In psychology and education, for example, understanding is sometimes identified with intuitive theories [33], mental models [34], core knowledge systems [35], or other structured representations, though not all internal structures must be representational [36].

Model-based approaches can successfully explain the contrast between EQUATION and LOOK-UP. Within a range of specified conditions (e.g., cases contained within LOOK-UP and for which friction is negligible), EQUATION and LOOK-UP can exhibit exactly the same performance, yet their internal properties vary. One thing EQUATION possesses, but LOOK-UP lacks, is an explicit representation of the dependence relationship that holds between time and distance (i.e., the continuous quadratic relationship dictated by a physical law that represents the role of gravity).

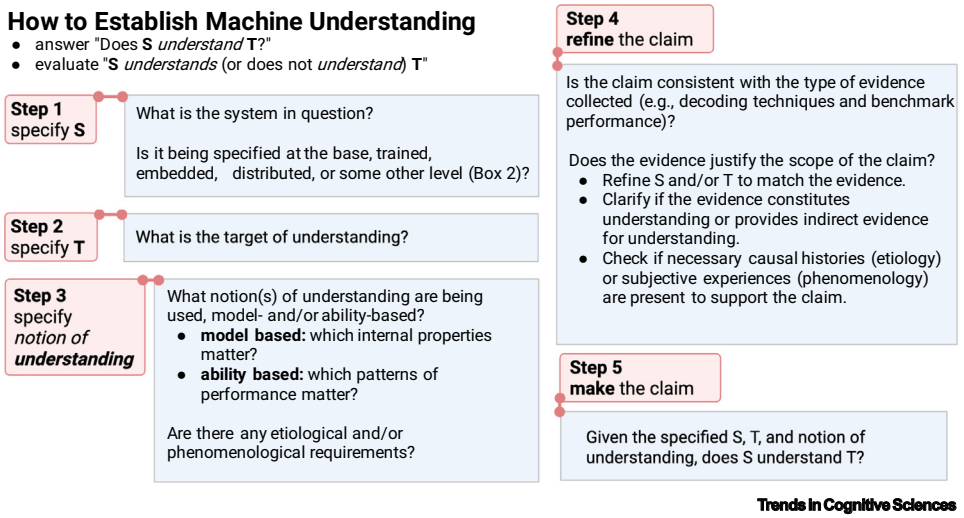

Figure 2. A five-step procedure for evaluating claims about a system's understanding of a target. This figure offers a five-step process for applying the framework summarized in Figure 1 to evaluate claims about machine understanding. Step 1 specifies the system S at the relevant level of analysis, for example, as a base, trained, embedded, or distributed system. Step 2 specifies the target T, since a system can understand one aspect of a domain but not another. Step 3 identifies the notion of understanding in question: model based, ability based, or a hybrid view, together with any etiological or phenomenological requirements. Step 4 refines the claim by checking whether the available evidence matches the selected notion of understanding and the scope of the claim. This includes distinguishing evidence that constitutes understanding from evidence that only supports an inference to understanding and potentially narrowing S or T when the evidence is more limited than the original claim suggests. This step highlights that the process of framing and evaluating claims about understanding can be iterative. Finally, Step 5 involves making a claim about whether S does or does not understand T.

A major challenge for model-based accounts is defining *which* internal properties constitute understanding and establishing reliable methods for evaluating those properties. For instance, many techniques aim to assess a system's world model [37,38] or internal representations [39], but these methods apply under limited conditions. Often, inferences concerning internal representations depend on patterns of performance or correlation, but such evidence is at best indirect, and these techniques can be unreliable [40,41]. Future work is needed—both theoretical and technical—to address these concerns.

Another common concern for model-based approaches is that having the right internal properties is not enough—those internal properties would not constitute understanding if they are causally ineffective. In the human case, for example, we might be reluctant to attribute an understanding of Newton's law of gravitation to someone who commonly misapplies it, even if research reveals that their brain encodes information that corresponds to Newton's law. This is one motivation for ability-based accounts, which we turn to next.

Ability-based accounts

On ability-based accounts, a system's understanding is constituted by its performance or behaviors, not the internal features that give rise to them. Ability-based accounts fit the widespread practice of evaluating AI systems based on their performance—for instance, applying benchmarks for natural language [42] or scene understanding [43] as a basis for attributing understanding of some target, such as language or scenes. Within philosophy, different ability-based accounts focus on different patterns of performance as key to understanding, such as identifying connections [44], answering counterfactual questions [45], making predictions

Box 3. World models in model-based understanding

Contemporary efforts to evaluate understanding in AI systems often appeal to the idea of a system's "world model", which Yildirim and Paul [154] define as "structure-preserving, behaviorally efficacious representations of the entities, relations, and processes in the real world". For example, a world model corresponding to disease transmission in a population could be a representation that preserves the spatial and causal relationships among individuals in that population. World models originated in cognitive science as a way to characterize representations in humans and nonhuman animals (see [155] for relevant discussion), but the idea has been widely adopted in AI, despite heterogeneity in how world models are defined and assessed.

What distinguishes world models from prior generations of AI models and representations? For many computer scientists, it is the requirement of structural correspondence to the relevant aspects of the world. When the term is used more loosely, it is often to highlight a model's impressive predictive power. For example, a world model can enable a robot to accurately forecast the consequences of its actions (e.g., what happens if a marble is rolled on a table) [155]. Similarly, a world model can directly train an AI agent by predicting and simulating the environment [21].

Do systems with world models possess understanding? The answer depends on both the target of understanding, T, and on one's account of understanding (see Figure 1). In practice, computer scientists use evaluations of world models to both attribute and deny understanding. As one example, an AI language model trained on legal board sequences for the game of Othello successfully recovered the underlying rules and states of the game (as assessed by probes) [37] (see also [156,157,158]), offering evidence for model-based understanding of the world of Othello. As another example [38], a system trained on taxi routes performed well on route prediction, but its underlying world model (as assessed by sequences of predictions) contained impossible street configurations, challenging an attribution of (model-based) understanding.

Whether world models developed in specific (even potentially broad) domains contain the requisite properties for deep or human-level understanding remains an area of active debate [6]. To what extent must the model be abstract, accurate, and coherent? Can it be isolated, or must it be integrated with more general knowledge? Is it enough to be "behaviorally efficacious" in a narrow set of tasks, or must it be flexible and generative? Does the requirement for behavioral efficacy sneak in ability-based commitments? These questions, which also arise for human cognition, are newly urgent in the context of AI.

[46], solving problems [47], recognizing instances of the target of understanding [48], constructing (scientific) models [46], or evaluating competing explanations [49].

Ability-based accounts have three well-known precedents in philosophy and cognitive science. First, philosopher Ludwig Wittgenstein's idea that the *meaning* of a word or expression can be identified with how it is used in a language—for short, that *meaning is use* [50]—has been extended to the claim that *understanding* is use [51]: one understands the word "bottle" when one can use it appropriately in different contexts. Since Wittgenstein denied that using a word or expression appropriately requires having a particular sort of representation "in the head," this can be seen as a rejection of a model-based approach in favor of an ability-based approach. Second, some forms of behaviorism claim that mental states refer to behavioral tendencies, not to internal properties, and so are constituted by performance or behavior [52]. Finally, the Turing test [53], in which an observer attempts to differentiate a machine from a human based on text exchanges, offers a performance-based criterion for intelligence; extending the approach to what constitutes understanding (vs. assessing intelligence) would yield an ability-based account of understanding.

Despite the ubiquity of benchmarks in AI evaluation, the way such benchmarks are actually used in practice suggests a relatively nuanced role for performance in assessing understanding. First, researchers typically attribute understanding to some target T that extends beyond the benchmark itself. For example, based on performance on a range of abstract reasoning and counterfactual tasks, a researcher might attribute (or deny) an ability for general reasoning [54,55], not understanding of the finite set of problems explicitly included in the benchmark. This suggests that observed performance is taken as evidence for the presence or absence of

some more general ability, rather than fully constituting understanding on its own. Second, AI practitioners often design and extend benchmarks or other measures of performance with an eye toward assessing *generalization*—that is, assessing the system's performance in cases beyond those on which the system was explicitly trained.

To understand this role of generalization, first consider assessments that employ independent and identically distributed test sets, in which training and test sets are assumed to be sampled from the same target distribution. In such cases, success on held-out or rare cases reflects generalization beyond the training set, but not necessarily out-of-distribution (OOD) generalization. However, contemporary AI evaluation often goes further. Many benchmarks for foundation models are deliberately designed to test adversarial, counterfactual, or compositionally novel cases that are likely to lie outside naturally occurring data distributions [56], and a failure to generalize is often treated as a failure of understanding (e.g., misclassifying familiar objects when presented from unusual perspectives [57]). This role of generalization again suggests that while ability-based practices are common, the link between benchmark performance and some target of understanding can be complex; benchmark performance is not always taken to constitute understanding but is instead treated as evidence for *something else*—either a further and often more general ability (on an ability-based account) or some internal characteristics (on a model-based account).

Turning to EQUATION and LOOK-UP, recall that the two systems can produce the same behavior under some conditions. However, a generalization task—predicting the distance for a novel time point (i.e., one that is not contained within LOOK-UP)—will reveal that EQUATION succeeds while LOOK-UP fails. By relying on generalization performance, an ability-based account of understanding can, therefore, deliver the verdict that EQUATION possesses understanding that LOOK-UP does not.

A major challenge for ability-based approaches is determining how to go from observed performance to the corresponding ability. For example, having observed that EQUATION performs well when predicting distance in a vacuum, is it appropriate to attribute to EQUATION some understanding of falling objects *in general*, or something more specific (such as medium-sized objects falling in a vacuum)? Further tests of performance can fine-tune attributions of understanding, but attributing abilities will typically go beyond previously observed performance. This is one reason researchers might look "inside" a system to its internal properties (as in a model-based approach)—doing so can provide a basis for going beyond a finite, observed set of behaviors to some more general ability.

A related challenge is that performance can be misleading. Performance can mask competence that would manifest under different conditions [58] or outstrip genuine ability. Even top-performing models overfit to commonly used benchmarks and struggle in tests with minor data distribution shifts [59] (see also [60,61]). AI researchers sometimes cite Goodhart's law, the adage that "when a measure becomes a target, it ceases to be a good measure" [62]; in our terms, performance on some measure becomes poor evidence for understanding when a system is optimized for that measure.

Etiology-based requirements

Etiology-based requirements are potential add-ons to model- or ability-based approaches. The intuition behind them is that it matters how an internal property or ability came about: it must have the right causal history. More specifically, the model's internal properties or abilities must have been appropriately caused by or connected to T (the target of understanding).

Consider an ImageNet-trained system that classifies a picture as “class 949” [63]. What (if anything) does the system understand? On a model-based approach, one might use decoding techniques to identify which image features are associated with class 949. On an ability-based approach, one might assess classification performance: which images does the system classify as 949? On an etiology-based view, the place to look is the system’s history: What are the images used to train class 949 images of? (Strawberries.) What were human coders referring to when labeling these images? (Strawberries.) It is answers to questions like these that, in conjunction with models or abilities, determine what the system understands.

Our use of the term “etiology” is borrowed from philosophy, where it means cause or origin, and refers to accounts that require an appropriate history. For example, etiological accounts of linguistic reference claim that terms acquire their meaning via a causal-historical chain of events: a person’s use of the word “Napoleon” refers to the historical Napoleon because that person is part of a community of language users that connects back to Napoleon [64] (see also [65–67]). Similarly, someone’s concept of “strawberry” refers to strawberries due to their direct or indirect causal contact with strawberries. In some sense, our “Napoleon/strawberry representations” and “Napoleon/strawberry abilities” were caused by the historical Napoleon and by actual strawberries. Other etiological approaches go beyond linguistic reference to representational content more broadly—on teleological approaches to mental content, for example, whether a mental state represents some target depends on its history—specifically, on whether it was subject to a selection process (such as natural selection or learning) in virtue of which it has the function of representing that target [68,69].

An etiology-based approach to understanding has received the most explicit defense in the context of language understanding in LLMs [9,70–72]. Some argue that LLMs can successfully refer to Napoleon and strawberries because they are causally connected to the humans who wrote the texts that the systems are trained on, and at least some of those humans were appropriately connected to Napoleon and strawberries. Critics of this view argue that establishing an appropriate connection to the world requires sensorimotor grounding [73] or communicative intentions that LLMs lack [74]; but see [75–77].

While etiological approaches have not (to our knowledge) been developed as general accounts of understanding, there are additional contexts in which they might be appealing, such as AI personalization [78]. Suppose an AI therapist learns about a patient, Alex, from a history of interactions with Alex. Suppose further that interactions with a different Alex (we willl call her Alex B.) would have resulted in exactly the same current internal state for the system (because Alex and Alex B. are similar in relevant respects), but the system has never interacted with Alex B. If we are tempted to say that the AI therapist understands Alex, but not Alex B., this reflects the system’s etiology (see [79] for a philosophical precedent to this example, Hilary Putnam’s Twin Earth).

Etiology is related to two other ideas: grounding and embodiment. Philosophers and psychologists talk about the grounding of symbols in perceptual experience [80], and roboticists discuss the grounding of language in action and perception [81]. However, the term “grounding” is sometimes used more broadly than etiology (e.g., to include connections between symbols or modalities). “Embodiment” usually involves some role for the body and its interactions with the world [82,83], which can serve as a basis for establishing the relevant etiology for understanding. For example, systems such as Toolformer [84] and ReAct [85] interface

with external tools, providing direct contact with the world. Humans can also act as a bridge to the world by providing supervised annotations [63], noisy image captions [86], or explicit human feedback [87]. In such cases, humans introduce a causal pathway that links the world to model outputs [88] and connects the humans' own etiology to that of the models [89].

In some cases, the target of understanding is not the external world; it can involve virtual environments where agents interact and learn. For instance, embodied agents can learn object names and apply this knowledge for instruction execution in 3D environments [90], while symbolic communication can emerge among agents with direct contact with their virtual worlds [91–93]. In these cases, etiology-based accounts would require that the system have an appropriate form of contact with a virtual entity, rather than the external world, in order for the system to understand that entity in the virtual world.

Etiology-based views can explain the contrast between EQUATION and LOOK-UP, but some nuance is required. If LOOK-UP's entries correspond to actual empirical measurements, it possesses a causal history directly tied to specific real-world objects and settings (such as the effects of air friction on Earth). In this sense, LOOK-UP is grounded in high-fidelity real-world data that might license an attribution of understanding, but the target of understanding might be narrow: due to its internal properties and abilities, it seems too generous to attribute understanding of *gravity* or *Newtonian laws*. In the case of EQUATION, nothing was specified about the system's causal history, but its internal properties and abilities provide good evidence that it, too, was in some way shaped by real-world measurements, and its representations and abilities potentially license broader attributions of understanding; for instance, of gravity or Newtonian laws. Thus, judgments about EQUATION versus LOOK-UP can diverge, on an etiological basis, if their internal properties or performance are grounded in different causal histories, and diverging intuitions about the two cases could reflect different assumptions about their etiologies.

Phenomenology-based requirements

The intuition behind phenomenology-based requirements is that understanding sometimes depends on having certain feelings or subjective experiences [94]. Consider the philosopher Stephen Turner's claim that "When a mother tells her 13-year-old daughter that she does not know what 'love' is, she is not making a comment about semantics; she is pointing to the nonlinguistic experiential conditions that are bound up with the understanding of the term that the daughter does not share" [95] (see also [96]).

On some ways of developing phenomenology-based requirements, they are a special instance of a model-based approach (i.e., one in which the relevant internal representation is of a subjective experience) or an etiological requirement (i.e., one in which the relevant causal antecedent to some representation or ability is a subjective experience).

Some evidence suggests that people attribute phenomenology to AI systems [97], and that such attributions matter for attributions of understanding. For instance, the "artificial empathy paradox" [98] describes how AI-generated empathy loses its perceived value once users recognize it is artificial. However, most often, when phenomenology arises in discussions of AI, it is to reject a phenomenology-based view [99,100], or to deny machine understanding (i.e., on the grounds that AI systems lack phenomenology) [101]. In the case of humans, some point to common mismatches between phenomenology and understanding; for instance, the *sense* of understanding is not a reliable guide to *actual*

understanding, which challenges at least one kind of phenomenology as a basis for understanding [102,103].

Within philosophy and cognitive science, subjective experience is sometimes highlighted in the context of interpersonal understanding, which seems to require reconstructing others' perspectives "from within", engaging with their viewpoint on its own terms [104]. Take empathetic understanding [105]: to truly understand another person experiencing grief, it is plausible that we need to successfully put ourselves in their shoes and that this requires a phenomenological experience of what grief is like.

A phenomenology-based approach does not offer a ready diagnosis of what differentiates EQUATION from LOOK-UP. However, returning to the Chinese Room, which is somewhat analogous, it is plausible that a Chinese speaker (implementing something like EQUATION) experiences a distinct phenomenology from the person in the Chinese room (implementing something more like LOOK-UP), and that this difference contributes to why we are inclined to attribute understanding in the former case and not the latter.

There are currently no widely accepted accounts of phenomenology or its measurement [94,106], making it difficult to explore this direction. Even from a perspective sympathetic to the role of phenomenology in understanding, phenomenology-based considerations will plausibly identify necessary causal antecedents to understanding (rather than constituting understanding). For example, it could be that the only way to construct an internal model for love is to experience love, but the experience itself does not constitute understanding; the resulting representation does.

Hybrid accounts

Our accounts of understanding support graded understanding (see Box 4) and can be combined. As noted already, etiological and phenomenological requirements are additions to model- or ability-based approaches. However, models and abilities can also be incorporated into a single account. For example, Wilkenfeld's "Understanding as compression" [107] integrates representational elements (model based) with functional capabilities (ability based) (see also 108).

In some cases accounts can be in tension. For example, the use of machine-generated synthetic data illustrates potential trade-offs between ability and etiology. While synthetic data can improve a system's performance [109], it weakens the etiological link to real-world phenomena. Conversely, the accounts can align, as in "model collapse", where recursively training models on synthetic output lead to irreversible failure in performance because the causal connection to the original data distribution is lost [110]. Charting the theoretical and empirical relationships across accounts is an important direction for future work (see Outstanding questions).

Implications for AI and beyond

Assessing machine understanding

Having offered a landscape of views about understanding, it might seem attractive to develop corresponding benchmarks for each view. Unfortunately, doing so would miss the point: benchmarks will reflect performance that could only constitute understanding on an ability-based account. On model-based accounts, benchmarks can only offer indirect evidence (for internal properties). The distinction between constitution and evidence is crucial to resolving otherwise unproductive debates, as in our introductory example with Drs Benchmark and Decode. We can now consider a more encompassing example. Imagine four (hypothetical)

Box 4. Deep versus shallow understanding

Understanding can come in degrees, and our accounts have implications for what differentiates shallow from deep understanding.

On model-based views, "deeper" understanding corresponds to "deeper" internal properties. For instance, on one view with roots in philosopher Philip Kitcher's work, one model is deeper than another if it includes more basic principles (e.g., a model that includes Newton's laws of motion will be deeper than a model that includes only Kepler's laws of planetary motion because the latter can be derived from the former) [159].

On ability-based views, "deeper" understanding corresponds to better or broader performance. For instance, world models can support deeper understanding to the extent they enable a broader range of successful behaviors (e.g., performing reliably even in OOD settings [57,124], or answering counterfactual questions in addition to conditional questions [54,56,160]).

With etiology-based requirements, deeper or better understanding can correspond to more numerous, direct, or accurate historical connections to the target of understanding. For example, a system that receives visual and verbal input to interpret natural language descriptions of physical objects [113] plausibly has a deeper understanding of strawberries than one that receives only verbal input, since it has more opportunities for causal influence from the target of understanding (strawberries).

Finally, with phenomenology-based requirements, deeper or better understanding could correspond to the breadth or intensity of subjective experience. For example, a person who has both linguistic and gustatory experience with strawberries might have a deeper understanding of strawberries than someone with only linguistic experience because they have richer and more diverse phenomenology. If we grant some phenomenology to AI models, we would expect a similar differentiation between text-only and multimodal LLMs.

Current directions in AI are pushing the boundaries of graded understanding, especially when it comes to etiology. Consider recent model-training methods, such as knowledge distillation, in which pre-trained "teacher" models are used to train smaller "student" models [161]. Do teacher and student models possess different levels of understanding by virtue of their differently mediated connections to the world? As a second example, some systems rely on lower-fidelity data for training but subsequently introduce "linear projection" layers or fusion techniques to generate output equivalent to that from systems trained with higher fidelity data (e.g., [162]). Even if two systems generate the same performance, does one system possess greater understanding by virtue of its grounding in higher fidelity data?

scientists debating whether a given AI system understands anxiety. Dr M's evidence comes from internal representations: a decoding technique has revealed intermediate layers corresponding to key features of anxiety. Dr A's evidence comes from performance: the system shows excellent performance on key benchmarks. Dr E's evidence comes from the system's history: its training and interactions in the world. And Dr P's evidence comes from phenomenology—an account of subjective experience and how the system does (or does not) experience anxiety. Whose evidence should take precedence, especially when the scientists reach conflicting conclusions? The answer is that it depends on one's account of understanding, and if that is not made explicit and introduced into the debate, these scientists will simply talk past one another.

This example might seem like no more than a thought experiment, but analogous cases have deep implications for downstream applications. For instance, whether it is deemed acceptable (legally, ethically, and medically) to deploy an AI therapist chatbot could depend precisely on what relevant stakeholders take the system to understand. Or, and perhaps more controversially, consider whether it's acceptable to deploy an AI psychiatrist empowered to prescribe medications to treat anxiety. Again, judgments could diverge on the basis of different attributions of understanding, stemming not from a disagreement over the empirical facts but over how they bear on understanding. Similar arguments have played out in the AI fairness space, where conflicting definitions of what constitutes "fairness" have led to different conclusions and public debates[ii,iii]. Drawing on lessons from fairness, we anticipate that agreeing on what constitutes

understanding can help engineers, policymakers, and other stakeholders work together more proactively and constructively.

As a bonus, our accounts offer a taxonomy of forms of evidence that can be marshaled in assessing understanding. For example, model-based accounts highlight the value of model interpretability techniques [111]. Ability-based accounts align with current canonical benchmarking practices, such as evaluating generalization [43,63] and robustness [112]. Etiology-based accounts foreground multimodal metrics with strong causal connections to the target of understanding [113]. And finally, phenomenology-based accounts might be most relevant in the context of embedded systems with a human in the loop, where the person's phenomenology could play a role even if the AI system itself has no phenomenology. Taken together, such evidence offers a potential antidote to a major limitation in current AI evaluation practices: overreliance on singular or aggregate performance metrics [114].

Recognizing different forms of evidence has another benefit: it forces us to be explicit about the assumptions guiding inferences from evidence to understanding. Consider legal understanding. When law students take the bar exam, a high score is treated as evidence of the legal understanding necessary to be licensed to practice law (but see [15]). This could be because exam performance is deemed constitutive of legal understanding, but more commonly, exam scores are taken as good evidence of understanding—the scores plausibly reflect relevant content knowledge, along with relevant experiences and abilities (a mix of model-, ability-, and etiology-based considerations). GPT-4 can now score within the top 10% on the bar exam [115] and could perform better soon. However, this will not reflect any real-world legal experience and might not reflect the generalization ability that the test assesses in humans. Does it possess legal understanding?

Making an inference from some source of evidence (such as bar exam performance) to a claim about understanding requires some account of understanding (see also [116]). When that inferential step is spelled out, we can better assess whether it is justified. Failing to do so is risky: when the assumptions that underwrite inferences from performance to understanding in humans do not generalize to AI systems, the trust engendered by attributions of understanding could be misguided.

Advancing machine understanding

One reason to define machine understanding is to create a road map toward machine intelligence. Many accounts of machine intelligence appeal to understanding; for instance, suggesting that "the ultimate goal of AI research is to build machines that can understand the world around us" [117], see also [74,118].

To the extent that AI systems are human tools, what might matter most is performance, and understanding is just a means to that end [119]. This speaks in favor of ability-based accounts, which would align the aim of achieving machine understanding most directly with performance.

Yet not all AI systems are designed solely for their performance alone [120]. In some cases, the system itself can be an illuminating object of study, such as when we use AI systems to better understand winning strategies in chess [121], human cognition, or intelligence itself. For these purposes, we might favor model-based accounts.

A third role for AI systems is as social partners, such as therapists, teachers, or companions. For these purposes, we might care about interpersonal understanding, for which etiological and phenomenological components plausibly play a larger role. However, pursuing machine

understanding on a phenomenology-based account introduces ethical concerns that go beyond the scope of this paper—for instance, what our moral obligations are to sentient machines [122].

Which approach to understanding is most useful will depend on the goals of stakeholders and of a given AI system, and which account is adopted can have implications for whether machine understanding is a goal we ought to pursue. In short, clarifying machine understanding does not just have implications for how AI researchers should talk, but also for what they should do.

The broader significance of our framework

Our framework is not merely a theoretical exercise; it offers a road map for addressing the complexities of AI in practice. In this final section, we identify distinct areas of practical significance for AI practitioners and researchers, cognitive and social science researchers, and public engagement.

For AI practitioners and researchers, our framework provides a basis for developing more holistic approaches to advancing AI. Current practice for assessing AI systems centers on performance benchmarks, which aligns most closely with an ability-based account [42,43,123]. While useful, this singular focus can obscure other challenges that arise at all stages of the AI lifecycle (see Table 1). These challenges are often addressed in isolation, including data and annotation bias (a largely etiological challenge), the black box problem (a largely model-based challenge; see Box 5), and OOD robustness (an ability-based challenge). However, they can be powerfully recast as specific concerns about the understanding a system does or does not possess. By utilizing our framework, practitioners can recognize that many of AI's most pressing challenges are, at their core, questions about the nature of understanding.

One consequence of this unification is that disparate techniques within machine learning can be integrated under a common goal: assessing and advancing a system's understanding. Methods such as decoding to achieve interpretability (for black box issues), counterfactual task design (for generalization), and data curation (for fair representation) are often treated as independent tasks. However, with a holistic approach under the umbrella of understanding, we gain a more comprehensive picture that prevents the catastrophic failures often discovered only after real-world deployment[i] [13]. These real-world failures often arise when we mistake narrow behavior-based performance for genuine understanding. The pluralist application of our framework, described in Table 1 illustrates what a more multifaceted approach to AI can look like at every stage of the pipeline, from data curation and engineering to model interpretability and benchmarking, ultimately helping us advance safer and more responsible AI.

For cognitive and social science researchers, our framework offers a bridge for interdisciplinary collaboration. Complex AI challenges (such as OOD generalization [57,124] or model collapse from synthetic data [110]) are typically framed as technical problems, which obscures their relevance outside of engineering. Our framework reorients such challenges by linking them to understanding, a construct with deep roots in cognitive science, philosophy, and education. This creates a shared conceptual ground for leveraging methods and solutions across fields that already study "understanding" to inspire and address core challenges facing AI.

This cross-disciplinary intellectual pollination has also been advocated by researchers in nonengineering fields to improve the design and evaluation practices of AI. For example, current methods for studying AI behaviors have been inadequate because they often rely on biased,

Table 1. Sample questions to guide AI research

AI development stage	Relevant account of understanding	Sample questions for AI research
Data and resources Data acquisition; feature engineering; and selection of dynamic runtime resources	Etiology based	Does our data acquisition process ensure a faithful causal link between the data and the target of understanding (e.g., more natural and direct connections to the target)? How might our data curation choices (e.g., using synthetic data, human annotations, and knowledge distillation) alter the system's causal history with respect to the target of understanding? How do we make claims about the target of understanding while accounting for bias and errors (e.g., data selection and data availability bias, data representation gaps, human annotator social lens, and annotation ambiguity)?
Model design and training Designing model architectures; defining learning objectives and algorithms; and training and optimization	Model based	How can we design model architectures and learning objectives that encourage the formation of 'deeper' models, such as those that represent more fundamental principles? What internal representations (e.g., knowledge structures) or other properties (e.g., model architecture) should our system possess to best capture the target?
	Ability based	What specific range of abilities (e.g., answering counterfactual questions and demonstrating robustness to OOD data) are we targeting, and how does this inform our learning objective and optimization?
	Etiology based	How can we formulate a model architecture design to preserve the model's connection to the target of understanding (e.g., choosing between an early fusion architecture vs. a late fusion architecture in multimodal learning)? How can we use techniques (e.g., fine-tuning with human feedback, real-time annotations, and personalization) to incorporate real-world feedback loops and strengthen a model's causal connections to the target of understanding?
	Phenomenology based	Is there a possibility that a system can be designed to possess phenomenology? If so, how is the anticipated phenomenology relevant to the target of understanding? Is there a risk of creating a system that is subject to moral consideration, and if so, does that have implications for development?
System evaluation Evaluation design and benchmark development	Model based	Do our interpretability and decoding techniques confirm that the model is learning the intended representations and algorithms, rather than just surface-level pattern matching? How can we measure the fidelity of internal representations to ensure alignment with design purposes and to prevent safety-critical failures, such as hallucination? To what extent does the system's internal representation or world model accurately represent the target's true data-generating process?
	Ability based	In designing evaluation benchmarks, how can we meaningfully distinguish true generalization from overfitting to the training data (particularly with massive-scale and/or private training data)? How do we design a suite of rigorous tests, including OOD, adversarial, and counterfactual scenarios, to accurately probe the boundaries and limitations of the system's abilities and to avoid catastrophic failures? Can we ensure that the relative performance of different models under an evaluation protocol will accurately reflect each model's real-world competence?
	Etiology based	Does benchmark performance provide evidence that the intended causal connections to the target of understanding have been preserved?
	Phenomenology based	Is it possible to assess the system's experience? If so, how is it related to the target of understanding? What are the ethical implications of evidence that a system has subjective experience?
AI and society AI explainability; transparency; and trust	Model based	How can we decode the AI system's internal representations of a target to generate explanations for the target or for the system's inner workings for its end users?
	Ability based	How can a system's performance be evaluated in the societal context in which it will actually be deployed?
	Etiology based	Can the system's outputs be connected to its real-world targets of understanding?
	Phenomenology based	Given that people will sometimes attribute conscious experience to AI systems when this attribution is likely unwarranted, what can be done to foster more accurate attributions and to support appropriate trust? How can human interactions with AI systems be structured such that the *human* is likely to experience the phenomenology relevant to a target of understanding?

Sample questions to guide AI research based on our proposed accounts of understanding, categorized by stages of AI development and interaction. This illustrates one way to use our framework: rather than adopting a single account of understanding, researchers can entertain multiple notions, or different notions for different stages of the design process, with the goal of creating systems that support various forms of understanding. On this pluralist approach, which kinds of questions to prioritize will depend on the developer's goals for the system.

Box 5. Assessing human understanding of machines

Our framework applies not only to assessing machine understanding, but also to human understanding of machines. With the rise of deep learning, AI has increasingly faced a "black box problem", whereby AI systems can be opaque not only to naïve users but also to the scientists and engineers who create them [163]. Concerns about transparency, interpretability, and explainability are fundamentally about understanding [164], and each of our accounts suggests a corresponding set of tools for human understanding of machines.

On a model-based approach, the aim is to provide humans with the information or experiences necessary to construct a mental model of the relevant internal properties of an AI system. For example, researchers have provided insight into intermediate units in neural networks [111] or have extracted the "hidden knowledge" possessed by systems such as AlphaZero [39,121].

On an ability-based approach, the aim is to induce the right abilities in humans—for example, their ability to answer "what if things had been different?" questions about the system's performance. One example is reporting benchmark performance across a range of human-understandable settings [114]; another is developing techniques to help human users answer counterfactual questions about the AI's predictions on a given input [165,166]. Characterizing the resulting representations and properties of AI models under different training conditions, as well as similar investigations at the base systems level through accumulating evidence from trained systems, can be thought of as providing a hybrid of model-based and ability-based understanding.

Etiology-based approaches might aim to connect humans to the target of understanding in the right way, for example, by providing some of the data (or relevant features of the data) that the model was trained on [167,168]. These approaches could provide insight into biases in those data or reveal how the model could have learned to capture real-world constructs.

Finally, phenomenology-based approaches could aim to communicate the subjective experiences of AI systems to humans (though this remains, at present, in the realm of science fiction).

In pursuing efforts to improve explainability, transparency, or interpretability, stakeholders can follow a process similar to that depicted in Figure 2 in the main text but where the system in question is a human, and the target of understanding is an AI system at some level of specification (see Box 2). Different ways of specifying understanding within our framework will align with different approaches to engendering understanding in humans. Our four accounts thus provide a basis for more precisely characterizing not only machine understanding but also human understanding of machines.

idealized human baselines—a problem Cameron Buckner terms "anthropofabulation" [125]. Buckner argues that future evaluations could instead draw inspiration from comparative psychology, adopting rigorous, species-fair testing protocols that assess underlying cognitive capacities rather than just surface-level performance [125].

Finally, our framework offers a powerful tool for advancing AI literacy by introducing nuance to public debates over whether and what AI systems "understand". Public perception is currently shaped by polarized media narratives and direct interactions with AI systems, both of which can be deeply misleading. Media discourse often promotes dichotomies between AI as a villain (e.g., AI systems becoming more powerful than humans) or a hero (e.g., AI revolutionizing medicine to extend human life) [126,127]. Simultaneously, direct user interaction can be deceptive, masking a lack of genuine understanding with behaviors such as sycophancy — the tendency of an AI system to generate responses that flatter or align with the user's stated preferences, even when those preferences might be factually incorrect or inappropriate [128]. Both channels overemphasize surface-level behavioral performance and encourage a binary view that an AI system either does or does not understand. Our framework invites the public to go beyond yes/no questions about machine understanding to appreciate different forms of understanding, and in so doing, exposes the relevance of "hidden" issues, such as model collapse or knowledge distillation, which might otherwise seem like abstract technical details. Linking a system's internal properties, abilities, etiology, and phenomenology to understanding makes it clear what's at stake in how a system is created, assessed, and deployed.

Concluding remarks

Advancing and assessing machine understanding are crucial to two major projects in contemporary AI: creating intelligent systems and evaluating whether and when we can trust them. The aim of this feature review article has been to develop a conceptual framework for machine understanding (see Figure 1). Rather than arguing for a particular approach, we have charted a landscape of options that can help those working towards and with AI to ask more precise questions, make more precise claims, and work more productively towards sophisticated and responsible forms of machine intelligence.

Author contributions

All authors jointly developed the conceptual framework. H.C. took the lead in preparing the manuscript, wrote the first draft of the manuscript, and took the lead in preparing figures and tables. S.R.G. provided expertise on the philosophical accounts of understanding. O.R. contributed expertise on AI and machine learning. T.L. contributed expertise in cognitive science and supervised the project. All authors played a role in writing and revising the manuscript.

Acknowledgments

We are deeply grateful to a number of people for providing helpful feedback and discussion, either in person or in writing. This includes Mel Andrews, Florian Boge, Allison Chen, Henry Conklin, David Danks, Will Fleisher, Antske Fokkens, Tom Griffiths, Aaron Hertman, Kareem Khalifa, Been Kim, Sanmi Koyejo, Sarah-Jane Leslie, Melanie Mitchell, Sam McGrath, John Morrison, Andrew Nam, Katie O'Dell, Russ Poldrack, Annika Schuster, Rich Shiffrin, Brandon Stewart, Frauke Stoll, Michael Strevens, Anna Tsvetkov, Phil Walsh, Angelina Wang, Lily Weng, Daniel Wilkenfeld, Rene van Woudenberg, William Yang, and Tyler Zhu. We are also grateful to Kathryn McGregor for help with manuscript preparation and to the participants in the Varieties of Understanding conference at UNAM, the Princeton HCI seminar series, and the Princeton Visual AI reading group for their helpful questions and discussion.

H. C. was supported by the Presidential Postdoctoral Research Fellowship at Princeton University. This material is partially based on work supported by the National Science Foundation under grant no. 2145198. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We are also grateful to Natural and Artificial Minds, a research initiative within the Princeton Laboratory for Artificial Intelligence, for supporting this collaboration.

Competing interests

The authors declare no competing interests.

Resources

[i]https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/

[ii]https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story

[iii]https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

References

1. Sellars, W.S. (1963) Philosophy and the scientific image of man. In *Science, Perception, and Reality* (Colodny, R., ed.), pp. 35–78, Humanities Press/Ridgeview
2. Strevens, M. (2008) *Depth: An Account of Scientific Explanation*, Harvard University Press
3. Zagzebski, L.T. (1996) *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*, Cambridge University Press
4. Elgin, C.Z. (2017) *True Enough*, MIT Press
5. Kvanvig, J.L. (2003) *The Value of Knowledge and the Pursuit of Understanding*, Cambridge University Press
6. Mitchell, M. and Krakauer, D.C. (2023) The debate over understanding in AI's large language models. *Proc. Natl. Acad. Sci.* 120, e2221907120
7. Mitchell, M. (2023) AI's challenge of understanding the world. *Science* 382, eadm8175
8. Mahowald, K. *et al.* (2024) Dissociating language and thought in large language models. *Trends Cogn. Sci.* 28, 517–540
9. Mandelkern, M. and Linzen, T. (2024) Do language models' words refer? *Comput. Linguist.* 50, 1191–1200
10. Sejnowski, T.J. (2023) Large language models and the reverse Turing test. *Neural Comput.* 35, 309–342
11. Wang *et al.* (2023) Scientific discovery in the age of artificial intelligence. *Nature* 620, 47–60
12. Yiu, E. *et al.* (2024) Transmission versus truth, imitation versus innovation: What children can do that large language and language-and-vision models cannot (yet). *Perspect. Psychol. Sci.* 19, 874–883
13. Obermeyer, Z. *et al.* (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 447–453
14. Moor *et al.* (2023) Foundation models for generalist medical artificial intelligence. *Nature* 616, 259–265
15. Kapoor, S. *et al.* (2024) Promises and pitfalls of artificial intelligence for legal applications. *J. Cross-discip. Res. Comput. Law* 2, 1–12

Outstanding questions

Which account of machine understanding is most useful for efforts to improve or assess machine intelligence? How does this depend on the domain or dimension of intelligence being advanced?

Which account of machine understanding is most useful for efforts to improve or assess artificial intelligence safety or transparency? How does this depend on the domain or dimension of safety or transparency being advanced?

How do the accounts of machine understanding that we propose relate to each other? In particular, to what extent do ability-based evaluations (e.g., out-of-distribution generalization and adversarial robustness) depend on implicit commitments about representation and invariance that overlap with model-based accounts? If such evaluations require the evaluator to specify which features are relevant and what invariances should hold, does this mean that ability-based and model-based accounts converge in practice, even if they differ in principle? More broadly, what are the conditions under which the accounts naturally complement each other, and when might there be genuine trade-offs or tensions between them?

Can the accounts of machine understanding that we propose be fruitfully extended to understanding in humans and nonhuman animals? While the understanding of various targets will vary across systems (human, nonhuman animal, and machine), are there reasons to have different accounts of understanding across systems?

16. Holmes, W. and Tuomi, I. (2022) State of the art and practice in ai in education. *Eur. J. Educ.* 57, 542–570
17. Krathwohl, D. (2002) A revision Bloom's taxonomy: An overview. *Theory Pract.* 41, 212–218
18. Bender, E.M. *et al.* (2021) On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623
19. Searle, J.R. (1980) Minds, brains, and programs. *Behav. Brain Sci.* 3, 417–457
20. Block, N. (1981) Psychologism and behaviorism. *Philos. Rev.* 90, 5–43
21. Ha, D. and Schmidhuber, J. (2018) World models. *arXiv* preprint arXiv:1803.10122
22. Epstein, R.A. *et al.* (2017) The cognitive map in humans: Spatial navigation and beyond. *Nat. Neurosci.* 20, 1504–1513
23. Yildirim, I. *et al.* (2020) Physical object representations for perception and cognition. In *The Cognitive Neurosciences* (6th edition) (Gazzaniga, Mangun, ed.), The MIT Press
24. Jara-Ettinger, J. *et al.* (2016) The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends Cogn. Sci.* 20, 589–604
25. Kim, J. (2010) Reasons and the first person. In *Essays in the Metaphysics of Mind*, pp. 105–124, Oxford University Press
26. Speer, R. *et al.* (2017) ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence* (31)
27. Werner, L. *et al.* (2023) Knowledge enhanced graph neural networks. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10
28. Kim, J. (1994) Explanatory knowledge and metaphysical dependence. *Philos. Issues* 5, 51–69
29. Ichikawa, J.J. and Steup, M. (2024) The analysis of knowledge. In *The Stanford Encyclopedia of Philosophy (Metaphysics Research Lab, Stanford University, 2024)* (Fall 2024 edn) (Zalta, E.N. and Nodelman, U., eds)
30. Shanahan, M. *et al.* (2023) Role play with large language models. *Nature* 623, 493–498
31. Ma, W. and Valton, V. (2024) Toward an ethics of AI belief. *Philos. Technol.* 37, 76
32. Herrmann, D.A. and Levinstein, B.A. (2024) Standards for belief representations in LLMs. *Mind. Mach.* 35, 5
33. Gopnik, A. and Meltzoff, A.N. (1997) *Words, Thoughts, and Theories*, MIT Press
34. Gentner, D. and Stevens, A.L. (2014) *Mental Models*, Psychology Press
35. Carey, S. and Spelke, E. (1996) Science and core knowledge. *Philos. Sci.* 63, 515–533
36. Thelen, E. and Smith, L.B. (1994) *A Dynamic Systems Approach to the Development of Cognition and Action*, MIT Press
37. Li, K. *et al.* (2023) *Emergent world representations: Exploring a sequence model trained on a synthetic task*, International Conference on Learning Representations
38. Vafa, K. *et al.* (2024) Evaluating the world model implicit in a generative model. *Adv. Neural Inf. Proces. Syst.* 37, 26941–26975
39. McGrath *et al.* (2022) Acquisition of chess knowledge in AlphaZero. *Proc. Natl. Acad. Sci.* 119, e2206625119
40. Bilodeau, B. *et al.* (2024) Impossibility theorems for feature attribution. *Proc. Natl. Acad. Sci.* 121, e2304406120
41. Ramaswamy, V.V. *et al.* (2023) Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10932–10941
42. Wang, A. *et al.* (2019) SuperGLUE: a stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems* (Vol. 32) (Wallach, H. *et al.*, eds), pp. 3261–3275, Publisher is Curran Associates, Inc.
43. Heilbron, F.C. *et al.* (2015) ActivityNet: A largescale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970
44. Kvanvig, J.L. (2018) Knowledge, understanding, and reasons for belief. In *The Oxford Handbook of Reasons and Normativity* (Star, D., ed.), pp. 685–704, Oxford University Press
45. Hills, A. (2016) Understanding why. *Noûs* 50, 661–688
46. de Regt, H.W. (2015) Scientific understanding: Truth or dare? *Synthese* 192, 3781–3797
47. Sullivan, E. and Khalifa, K. (2019) Idealizations and understanding: Much ado about nothing? *Australas. J. Philos.* 97, 673–689
48. Strevens, M. (2024) Grasp and scientific understanding: a recognition account. *Philosophical Studies* 181, 741–762
49. Newman, M. (2016) An evidentialist account of explanatory understanding. In *Explaining Understanding* (Grimm, S.R. *et al.*, eds), pp. 190–211, Routledge
50. Wittgenstein, L. (2009/1953) *Philosophical investigations* (revised 4th edn.), Wiley-Blackwell
51. McGinn, C. (1984) *Wittgenstein on meaning: An interpretation and evaluation*, Basil Blackwell
52. Skinner, B.F. (2016) Why I am not a cognitive psychologist. In *Approaches to Cognition: Contrasts and Controversies* (Knapp, T.J. and Robertson, L.C., eds), pp. 79–90, Routledge
53. Turing, A.M. (2007) Computing machinery and intelligence. In *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer* (Epstein, R. *et al.*, eds), pp. 23–65, Springer
54. Pearl, J. (2019) The seven tools of causal inference, with reflections on machine learning. *Commun. ACM* 62, 54–60
55. Moskvichev, A.K. *et al.* (2023) The ConceptARC benchmark: Evaluating understanding and generalization in the arc domain. *Transact. Mach. Learn. Res.*
56. Wu *et al.* (2024) Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Duh, K. and Gomez, H. and Bethard, A., eds), pp. 1819–1862
57. Alcorn, M.A. *et al.* (2019) Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4845–4854
58. Firestone, C. (2020) Performance vs. competence in human-machine comparisons. *Proc. Natl. Acad. Sci.* 117, 26562–26571
59. Recht, B. *et al.* (2019) Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine. Learning*
60. Srivastava, A. *et al.* (2022) Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv* preprint arXiv:2206.04615
61. Thomas, R. and Uminsky, D. (2020) The problem with metrics is a fundamental problem for AI. *arXiv* preprint arXiv:2002.08512
62. Strathern, M. (1997) 'Improving ratings': Audit in the British university system. *Eur. Rev.* 5, 305–321
63. Russakovsky, O. *et al.* (2015) ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252
64. Michaelson, E. (2024) Reference. In *The Stanford Encyclopedia of Philosophy (Fall 2024 Edition)* (Zalta, Edward N. and Nodelman, Uri, eds)
65. Geach, P.T. (1969) The perils of Pauline. *Rev. Metaphys.* 23, 287–300
66. Donnellan, K.S. (1970) Proper names and identifying descriptions. *Synthese* 21, 335–358
67. Kripke, S.A. (1972) *Naming and Necessity*, Harvard University Press
68. Schulte, P. (2022) Teleological Theories of Mental Content. In *The Stanford Encyclopedia of Philosophy (Summer 2022 Edition)* (Neander, K. and Zalta, E.N., eds)
69. Millikan, R.G. (2021) Neuroscience and teleosemantics. *Synthese* 199, 2457–2465
70. Butlin, P. (2023) Sharing our concepts with machines. *Erkenntnis* 88, 219–242
71. Søgaard, A. (2022) Understanding models understanding language. *Synthese* 200, 443
72. Koch, S. (2025) Babbling stochastic parrots? A Kripkean argument for reference in large language models. *Philos. AI* 1, 19–33
73. Harnad, S. (2024) Language writ large: LLMs, ChatGPT, grounding, meaning and understanding. *arXiv* preprint arXiv: 2402.02243

74. Bender, E.M. and Koller, A. (2020) Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Jurafsky, D. and Chai, J. and Schluter, N. and Tetreault, J., eds), pp. 5185–5198
75. Pepp, J. (2025) Reference without intentions in large language models. *Inquiry*, 1–19 https://doi.org/10.1080/0020174X.2024.2448482 Published online 9 January 2025
76. Lederman, H. and Mahowald, K. (2024) Are language models more like libraries or like librarians? Bibliotechnism, the novel reference problem, and the attitudes of LLMs. *Trans. Assoc. Comput. Linguist.* 12, 1087–1103
77. Attah, N.O. (2025) Do language models lack communicative intentions? *Synthese* 205, 187
78. Kirk, H.R. *et al.* (2024) The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nat. Mach. Intell.* 1–10
79. Putnam, H. (1975) The meaning of 'Meaning'. In *Minnesota Studies in the Philosophy of Science, Vol. VII: Language, Mind, and Knowledge* (Gunderson, K., ed.), pp. 131–193, University of Minnesota Press
80. Harnad, S. (1990) The symbol grounding problem. *Phys. D: Nonlinear Phenom.* 42, 335–346
81. Cangelosi, A. (2010) Grounding language in action and perception: From cognitive agents to humanoid robots. *Phys Life Rev* 7, 139–151
82. Pfeifer, R. and Bongard, J. (2006) *How the body shapes the way we think: A new view of intelligence*, MIT Press
83. Chrisley, R. (2003) Embodied artificial intelligence. *Artif. Intell.* 149, 131–150
84. Schick, T. *et al.* (2024) Toolformer: Language models can teach themselves to use tools. *Adv. Neural Inf. Proces. Syst.* 36
85. Yao, S. *et al.* (2023) ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*
86. Sun, C. *et al.* (2017) Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 843–852
87. Ouyang, L. *et al.* (2022) Training language models to follow instructions with human feed- back. In *Advances in Neural Information Processing Systems* (35) (Koyejo, S. *et al.*, eds)
88. Mollo, D.C. and Millière, R. (2023) The vector grounding problem. *arXiv* preprint arXiv:2304.01481
89. Denton, R. *et al.* (2021) Whose ground truth? Accounting for individual and collective identities underlying dataset annotation. *arXiv* preprint arXiv:2112.04554
90. Hill, F. *et al.* (2020) Grounded language learning fast and slow. *arXiv* preprint arXiv:2009.01719 (2020). 2009.01719
91. Mordatch, I. and Abbeel, P. (2018) Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the AAAI conference on Artificial Intelligence* (32)
92. Havrylov, S. and Titov, I. (2017) Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *Adv. Neural Inf. Proces. Syst.* 30
93. Lazaridou, A. *et al.* (2022) Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations*
94. Kriegel, U. (2015) *The Varieties of Consciousness*, Oxford University Press
95. Turner, S.P. (2018) *Cognitive Science and the Social: A Primer*, Routledge
96. Nagel, T. What is it like to be a bat? *Philos. Rev.* 83, 435–450.
97. Colombatto, C. and Fleming, S.M. (2024) Folk psychological attributions of consciousness to large language models. *Neurosci. Conscious.* 1, niae013
98. Perry, A. (2023) AI will never convey the essence of human empathy. *Nat. Hum. Behav.* 7, 1808–1809
99. Ng, G.W. and Leung, W.C. (2020) Strong artificial intelligence and consciousness. *J. Artif. Intell. Conscious.* 7, 63–72
100. Chalmers, D.J. (2023) Could a large language model be conscious? *arXiv* preprint arXiv:2303.07103
101. Bayne, T. *et al.* (2024) Tests for consciousness in humans and beyond. *Trends Cogn. Sci.* 28, 454–466
102. De Regt, H.W. (2017) *Understanding scientific understanding*, Oxford University Press, New York
103. Trout, J.D. (2002) Scientific explanation and the sense of understanding. *Philos. Sci.* 69, 212–233
104. Ismael, J. (2017) Why (study) the humanities? In *Making Sense of the World: New Essays on the Philosophy of Understanding* (Grimm, S.R., ed.), pp. 177–193, Oxford University Press
105. Jaspers, K. (1997) *General psychopathology*, 2. JHU Press
106. Van Gulick, R. (2025) Consciousness. In *The Stanford Encyclopedia of Philosophy (Spring 2025 Edition)* (Zalta, E.N. and Nodelman, U., eds)
107. Wilkenfeld, D.A. (2019) Understanding as compression. *Philos. Stud.* 176, 2807–2831
108. Potochnik, A. (2017) *Idealization and the aims of science*, University of Chicago Press
109. Savage, N. (2023) Synthetic data could be better than real data. *Nat. Outlook: Robot. Artif. Intell.* https://doi.org/10.1038/d41586-023-01445-8
110. Shumailov, I. *et al.* (2024) AI models collapse when trained on recursively generated data. *Nature* 631, 755–759
111. Bau, D. *et al.* (2017) Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2282–2290
112. Hendrycks, D. and Dietterich, T. (2019) Benchmarking neural network robustness to common corruptions and perturbations. *arXiv* preprint arXiv:1903.12261. Appeared in ICLR 2019, 1903.12261
113. Matuszek, C. *et al.* (2012) A joint model of language and perception for grounded attribute learning. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1435–1442
114. Burnell, R. *et al.* (2023) Rethink reporting of evaluation results in AI. *Science* 380, 136–138
115. Achiam, J. *et al.* (2023) GPT-4 technical report. *arXiv* preprint arXiv:2303.08774
116. Mancoridis, M. *et al.* (2025) Potemkin understanding in large language models. In *Forty-second International Conference on Machine Learning*
117. Bengio, Y. *et al.* (2013) Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1798–1828
118. Lake, B.M. *et al.* (2017) Building machines that learn and think like people. *Behav. Brain Sci.* 40, e253
119. Legg, S. and Hutter, M. (2007) Universal intelligence: A definition of machine intelligence. *Mind. Mach.* 17, 391–444
120. Kurzweil, R. (1990) *The Age of Intelligent Machines*, MIT Press
121. Schut, L. *et al.* (2025) Bridging the human–AI knowledge gap through concept discovery and transfer in AlphaZero. *Proc. Natl. Acad. Sci.* 122, e2406675122
122. Gibert, M. and Martin, D. (2022) In search of the moral status of AI: Why sentience is a strong argument. *AI & Soc.* 37, 319–330
123. Cordts, M. *et al.* (2016) The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223
124. Hendrycks, D. *et al.* (2021) Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271
125. Buckner, C. (2019) The comparative psychology of artificial intelligences. *Philsci* Archive:16128
126. Cave, S. *et al.* (2019) "Scary robots" examining public responses to AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 331–337
127. Cave, S. and Dihal, K. (2019) Hopes and fears for intelligent machines in fiction and reality. *Nat. Mach. Intell.* 1, 74–78
128. Sharma, M. *et al.* (2023) Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*
129. Krenn, M. *et al.* (2022) On scientific understanding with artificial intelligence. *Nat. Rev. Phys.* 4, 761–769
130. Stueber, K.R. (2012) Understanding versus explanation? How to think about the distinction between the human and the natural sciences. *Inquiry* 55, 17–32
131. Grimm, S.R. (2016) How understanding people differs from understanding the natural world. *Philos. Issues* 26, 209–225
132. Grimm, S. (2024) Understanding. In *The Stanford Encyclopedia of Philosophy (Winter 2024 Edition)* (Zalta, E.N. and Nodelman, U., eds)
133. Moravcsik, J.M. (1979) Understanding and knowledge. In *Proceedings of the Aristotelian Society* (53), pp. 77–91

134. Zagzebski, L. (2019) Toward a theory of understanding. In *Varieties of Understanding: New Perspectives from Philosophy, Psychology, and Theology* (Grimm, S.R., ed.), pp. 123–137, Oxford University Press
135. Woodward, J. (2003) *Making things happen: A theory of causal explanation*, Oxford University Press
136. Lipton, P. (2001) *Inference to the Best Explanation* (2nd edn.), Routledge
137. Mără oiu, A.I. and Dumitru, M. (2024) *Understanding and conscious experience: Philosophical and scientific perspectives*, Routledge
138. Hempel, C.G. (1965) *Aspects of scientific explanation: And other essays in the philosophy of science*, Free Press
139. Bourget, D. (2017) The role of consciousness in grasping and understanding. *Philos. Phenomenol. Res.* 95, 285–318
140. McSweeney, M.M. (2023) Why Mary left her room. *Philos. Phenomenol. Res.* 109, 261–287
141. Baumberger, C. *et al.* (2017) What Is Understanding? In *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science* (Grimm, S.R. and Baumberger, C. and Ammon, S., eds), pp. 1–34, Routledge
142. Elgin, C.Z. (2009) Is Understanding Factive? In *Epistemic Value* (Pritchard, D. and Millar, A. and Haddock, A., eds), Oxford University Press
143. Khalifa, K. (2017) *Understanding, explanation, and scientific knowledge*, Cambridge University Press
144. Titus, L.M. (2024) Does ChatGPT have semantic understanding? A problem with the statistics-of-occurrence strategy. *Cogn. Syst. Res.* 83, 101174
145. Tamir, M. and Shech, E. (2023) Machine understanding and deep learning representation. *Synthese* 201, 51
146. Barman, K.G. *et al.* (2024) Towards a benchmark for scientific understanding in humans and machines. *Mind. Mach.* 34, 6
147. Beckmann, P. (2025) New horizons in machine understanding: explanatory and objectual understanding in deep learning video generation models. *Synthese* 206, 285
148. Raiaan, M.A.K. *et al.* (2024) *A review on large language models: Architectures, applications, taxonomies, open issues and challenges*, IEEE Access
149. Park, J.S. *et al.* (2023) Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual ACM symposium on user interface software and technology*, pp. 1–22
150. Ichter, B. *et al.* (2023) Do as I can, not as I say: Grounding language in robotic affordances. In *Proceedings of the 6th Conference on Robot Learning, vol. 205 of Proceedings of Machine Learning Research* (Arunkumar, A. and Boots, B. and Gordon, G. and Su, W., eds), pp. 287–318
151. Radford, A. *et al.* (2018) Improving language understanding by generative pre-training. https://openai.com/research/language-unsupervised
152. Wei, J. *et al.* (2022) Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Proces. Syst.* 35, 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abbf0-Abstract-Conference.html
153. Wang, L. and Ma, J. (2024) Multi-step chain-of-thought in geometry problem solving. In *2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS)*, pp. 1113–1117
154. Peeters, M.M. *et al.* (2021) Hybrid collective intelligence in a human–AI society. *AI & Soc.* 36, 217–238
155. Yildirim, I. and Paul, L. (2024) From task structures to world models: What do LLMs know? *Trends Cogn. Sci.* 28, 404–415
156. Li, X. *et al.* (2025) A comprehensive survey on world models for embodied AI. *arXiv* preprint arXiv:2510.16732
157. Toshniwal, S. *et al.* (2022) Chess as a testbed for language model state tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence* (volume 36), pp. 11385–11393
158. Nanda, N. *et al.* (2023) Emergent linear representations in world models of self-supervised sequence models. *arXiv* preprint arXiv:2309.00941
159. Kitcher, P. (1989) Explanatory unification and the causal structure of the world. In *Scientific Explanation, vol. XIII of Minnesota Studies in the Philosophy of Science* (Kitcher, P. and Salmon, W.C., eds), pp. 410–505, University of Minnesota Press
160. Lewis, M. and Mitchell, M. (2025) Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*
161. Hinton, G. *et al.* (2015) Distilling the knowledge in a neural network. *arXiv* preprint arXiv:1503.02531
162. Feng, C. *et al.* (2024) Naturally supervised 3D visual grounding with language-regularized concept learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13269–13278
163. Carvalho, D.V. *et al.* (2019) Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 832
164. Fleisher, W. (2022) Understanding, idealization, and explainable AI. *Episteme* 19, 534–560
165. Wachter, S. *et al.* (2017) Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31, 841
166. Ustun, B. *et al.* (2019) Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 10–19
167. Raji, I.D. *et al.* (2021) AI and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*
168. Meng, J. (2024) AI emerges as the frontier in behavioral science. *Proc. Natl. Acad. Sci.* 121, e2401336121.1