Viewpoints

Knowledge in Animals and Machines

L.A. Paul[1,2,3,4]

[1]Department of Philosophy, Yale University, New Haven, Connecticut 06511, [2]Department of Psychology, Yale University, New Haven, Connecticut 06511, [3]Wu-Tsai Institute, Yale University, New Haven, Connecticut 06511, and [4]Munich Center for Mathematical Philosophy, Ludwig Maximilian University of Munich, München 80539, Bavaria, Germany

Drawing on philosophical theories of knowledge, I develop a conceptual framework for knowledge in animals and machines. I use this framework to engineer a new concept of large language model (LLM) knowledge and propose a taxonomy of knowledge that spans biological and artificial categories, from C. elegans to LLMs. The approach lays a foundation for a rigorous account of how knowledge can be realized across animals and machines while at the same time respecting the diverse biological and physical sources of such knowledge.

Key words: **artificial intelligence; classification; concept; epistemology; instrumental; knowledge; LLM; machine; philosophy; representation; taxonomy; world model**

A large language model (LLM), the underlying machine basis for a "chatbot," such as Open AI's GPT series and Google's Claude, can write a philosophy paper that rivals an undergraduate thesis (Conroy, 2023). It can develop a conference program and summarize an academic manuscript (Luo et al., 2025). It can discuss your favorite jazz performance, drawing aesthetic conclusions and assessing nuance (Khadangi et al., 2025).

It can give you travel advice. If, say, you are an American interested in visiting gardens in Japan, you can ask a chatbot like ChatGPT to tell you about the Koraku-en Garden in Okayama Prefecture. Your chatbot can inform you that it is considered to be one of Japan's greatest gardens. It will tell you that Koraku-en is widely regarded as a masterpiece of Japanese landscape design and describe its key highlights. It might even come up with creative new ways to describe the garden or have ideas about how such gardens relate to other cultures. You can then follow up by asking for directions on how to get there from Tokyo using the best combination of bullet trains. In some sense, then, the LLM knows quite a bit about Koraku-en: likely more than you do.

But what is this sense? This is a way of asking what I term "the Knowledge Question":

What can an LLM know?

To answer the Knowledge Question, we need to take a systematic approach. I will begin by mapping out a preliminary taxonomy of knowledge in organisms. The plan is to leverage our implicit understanding of the neural grounds for knowledge in coordination with philosophical analysis to develop the conceptual framework that underpins our intuitive grasp of what knowledge requires. Once we have a preliminary taxonomy, we can extend the framework to include machines. The idea is to engineer a new concept of LLM knowledge using a taxonomy that can span categories from the biological to the artificial.

The approach lays a foundation for a rigorous account of how knowledge can be realized across a wide range of animals while at the same time respecting its diverse neurological and biological sources. It creates a structure that can guide productive debate and discussion, allowing us to articulate key questions at interdisciplinary borders. For example, what kinds and degrees of abilities to represent are realized by the varieties of neural systems we find in the animal kingdom? Which of these types of sensory inputs can support knowledge? What is the relationship between having a brain and having the capacity to represent the world? Last but not least, how are these questions related to emerging discussions of the nature and possibility of machine intelligence? A shared conceptual framework provides the scaffolding needed to ask these questions and to adjudicate disputes in an interdisciplinary context.

My hope is that my answer to the Knowledge Question and, more broadly, my conceptual approach to explicating machine knowledge will be of interest to neuroscientists who wish to ascribe capacities for knowledge, memory, and learning to diverse species. The approach should be of interest not just to cognitive neuroscientists and computational cognitive scientists but also to those who work in neuroscientific frameworks focused at the biological or cellular level. Recent advances in AI have opened this opportunity for neuroscientists, cognitive scientists, and philosophers to engage in an exploration of the conceptual foundations for the scientific study of knowledge. This *Viewpoint* constitutes a respectful attempt to start the conversation.

Taxonomy as Conceptual Framework

A taxonomy is a system of classification. It groups things into categories based on what kind of thing they are, and these kinds are

Received May 13, 2025; revised Nov. 13, 2025; accepted Nov. 15, 2025.

Author contributions: L.A.P. wrote the paper.

I am indebted to Daniel Colón-Ramos, Harvey Lederman, John Morrison, and Jonathan Schaffer for many helpful comments. I thank the members of the Cognition and Neural Computation Lab at Yale University for discussion of an early version of this work.

The author declares no competing financial interests.

Correspondence should be addressed to L.A. Paul at la.paul@yale.edu.

https://doi.org/10.1523/JNEUROSCI.0939-25.2025

in turn defined by similarities and differences at the intended level of description.

Such a classification gives us a conceptual framework for making sense of how two things can be very different and yet very similar at the same time. For example, I might have a dinner party where I offer my guests a glass of wine as well as a glass of water. Both liquids are potable and pleasant to drink; in this sense, they are the same kind of thing. At the same time, the dinner I've cooked relies on important differences between them. Water and wine have different underlying physical bases: they are composed of different types of molecules, resulting in different boiling points. In this sense, they are different kinds of things.

In our dinner party systematization of kinds, in the category of "Good things to drink," we can classify wine and water as the same kind of thing. Wine and water are both tasty, potable liquids that people like to drink at dinner parties and are liquids at room temperature that are socially acceptable to offer to dinner guests. Our classification reflects the uses we have for wine and water at a social level, that is, at this level of description.

Moving to the category "Things to eat," when we move to a lower level of description with respect to how such things are cooked, we classify wine and water as different kinds. At the thermodynamical level, water is good to cook with by boiling it at 100°C while wine is not. This difference in classification corresponds to differences in the average kinetic energy of the molecules that compose water versus those that compose wine. Wine boils at ~79°C, while water boils at 100°C, and a cook ignores this difference at their peril.

To cook the dinner successfully, it is essential that I understand the underlying differences between the two liquids, boiling the water for the potatoes while slowly simmering the wine reduction for the sauce. I do not need to know the details about the kinetic energies of wine molecules versus those of water molecules, but I do need to understand that wine and water evaporate at different temperatures. Figure 1 describes our dinner party taxonomy.

The dinner party taxonomy captures a way to understand different relationships between different kinds of foods, categorizing them at different levels of description depending on which properties are relevant. The way the space is carved up highlights relevant similarities of properties for items in one kind of context (e.g., a social context) and relevant differences between properties for items in another kind of context (e.g., cooking a sauce). Obviously, what belongs in what category admits of degree, and new culinary discoveries may add or rearrange the structure. The classification supports explanations (I kept the gas low on the reduction so as not to boil the wine) and predictions (I will need wine glasses as well as water glasses). It even supports novel hypotheses: if, perhaps, I mix water into my wine reduction, I should expect to turn the gas up to get it to simmer.

The Kingdom of Knowledge

We can use the taxonomic approach to systematize our understanding of knowledge of the external world. A taxonomy of knowledge of the external world will allow us to identify and classify kinds of properties that capture important, high-level physical features and behavioral similarities while distinguishing between lower-level systems that realize these properties.

To begin, start with some platitudes about how we apply the concept. We say things like "knowledge is power," "The more you know, the more you don't know," and "A little knowledge is a dangerous thing." Less cliché-ridden, we might say that knowledge requires information or that you can gain knowledge through education and experience. Consistent with this is the generalization that "Knowledge is the possession of information and the capacity to use that information to guide one's actions" (Stalnaker, 2012).

Moving past ascriptions of knowledge to people, we are also happy to say, for example, that dogs know who their owners are or that rats know where to find food in the city. In fact, we can apply the concept of knowledge to an even wider range of animals. We can say that *C. elegans* knows which temperature gradient it prefers and that an ant knows its way home (Luo et al., 2014; Schwarz et al., 2017). In all of these species, there is an important sense in which they possess information and use it to guide their actions. So, in accordance with our generalization of knowledge in terms of an entity's having the capacity to possess and use information to guide its actions, members of these species can know things.

Importantly, there is a common ground for these attributions of knowledge. All of these animals can be said to have knowledge

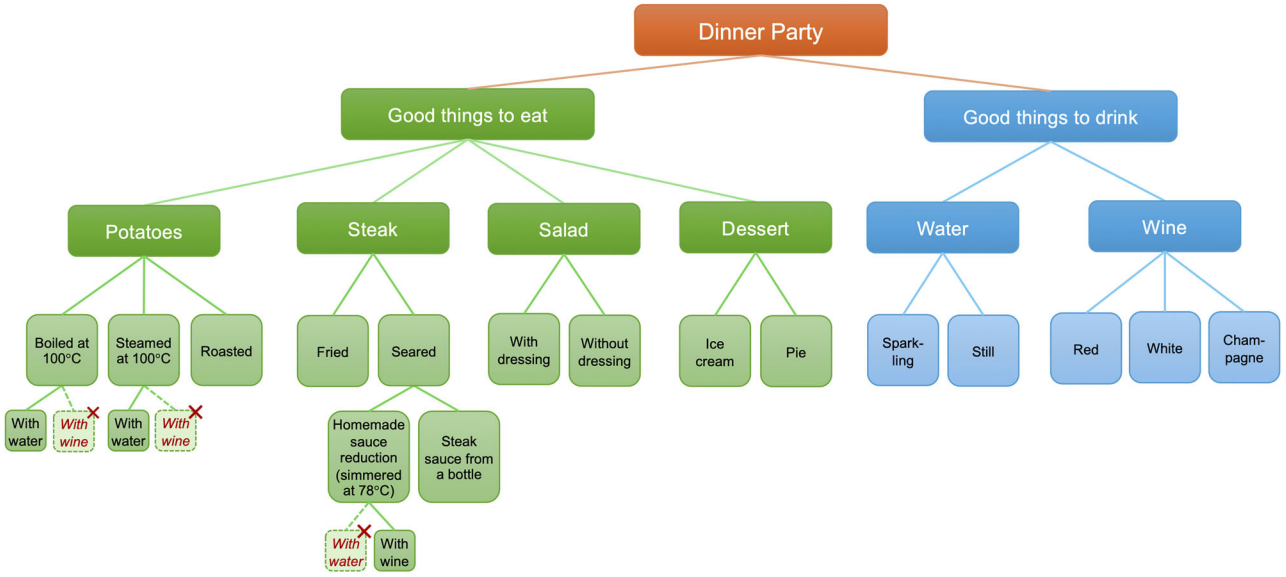

Figure 1. The dinner party taxonomy. Items in italics fail to belong to the classification; they are shown as rejected suggestions.

of the external world in virtue of the relationship between the external stimulus, their sensory inputs, and their neural response. There is a causal mapping from the external world, via inputs to its sensory apparatus, to the animal's neural system, and the animal's neural response defines its ability to guide its actions. There is also an implicit success condition. The knowledge results from an interaction with the external environment that represents or corresponds to the physical world in the right way. To ascribe knowledge, there must be enough "neurons to world" fit, such that the animal exhibits a sufficient level of behavioral flexibility and adaptive success in its goal-oriented behavior, sufficient to say their actions are being "guided."

At the same time, the biological structures that underpin these behaviors differ widely across different species. The most obvious contrast to animals like ants or *C. elegans* is human beings, with their huge brains and abilities to flexibly adapt to changes in their environments. From a computational perspective, most ordinary human knowledge of the external world relies on what computational cognitive scientists describe as "world models" for efficacious planning and successful action. "World models are structure-preserving, behaviorally efficacious representations of the entities and processes in the real world, including objects with 3D shapes and physical properties, scenes with topological relations and navigable surfaces, and agents with beliefs and desires. Human thought often relies on these types of world models to perceive, effectively reason, and talk about the world" (Yildirim and Paul, 2024a). World models represent objects and spaces, but they also represent and predict dynamical information such as the flow of water and the movement of living things, allowing agents to use them for flexible belief, learning, and action (Bi et al., 2024). These models are generated (in animals) by the brain through perceptual contact with the environment, and people use them to perform simulations of their environments that capture the causal structure needed for prediction, inference, and planning.

On this approach, brain processes and other neural activity in human beings generate such world models, which are characterized in terms of constituting mental representations that are caused by the right sort of interaction with the world. To say the representations are "structure preserving" is to say that the relevant structure of the world (e.g., objects with 3D shapes and physical properties, scenes with topological relations, and navigable surfaces) can be captured via a homomorphic mapping to the representation. More precisely, a representation is structure preserving when "the mapping from entities in the represented system to their symbols is such that functions defined on the represented entities are mirrored by functions of the same mathematical form between their corresponding symbols" (Gallistel and King, 2010; p. 55). We may also speak of approximate preservation of structure, allowing the notion of having a world model to be gradable.

Intuitively, to grasp the sense of "capturing structure," think of an architect's drawing of the interior of a room. The marks and arrangements on the drawing are representations of lengths and angles and other spatial relationships that capture the physical, spatial structure of the room and its contents. If it is a good drawing, it accurately represents this structure: there is a mapping from the room and its contents to the features of the drawing that preserves the relevant structural information. (We can relate this to Tolman's idea that knowledge can involve cognitive maps, as suggested by his sunburst experiment, where rats in a maze took a shortcut to a reward without previously having walked the route of the detour they chose; Tolman, 1948.)

We can contrast these neural systems and the richly representational world models they give rise to with the more minimal neural systems of animals like *C. elegans* and desert ants. In each case, the brain of the animal supports knowledge of the external world in virtue of its direct interaction with the world. Yet, it is unlikely that *C. elegans* and ants have brains that can support the type of rich representation needed for true world models. For example, in navigation, desert ants rely primarily on "dead reckoning," a mechanical task that involves the process of continuously evolving change of position rather than "piloting," which requires a richer type of representation (Gallistel, 1990).

We can distinguish between the way a human being can (ordinarily) know something about the external world in virtue of how that person represents and navigates the world via world models, versus the way that an animal like *C. elegans* can know something about the external world in virtue of how its neural activity allows it to flexibly navigate its local environment, without losing sight of the fact that in both cases the ultimate source of the knowledge stems from sensory contact with the environment. We can even use this similarity to speculate whether we have made a mistake in our taxonomic classification: perhaps the tiny brain of *C. elegans* realizes the smallest of world models.

Figure 2 presents an initial proposal for a taxonomy of knowledge based on relevant differences and similarities in the sensory systems of organisms. The proposal starts with an intuitively plausible distinction between organisms with brains and those without. Further distinctions make room for organisms that lack brains yet have neural systems that allow them to sense the environment in a way that gives them the capacity to flexibly navigate a local environment (e.g., jellyfish; Bielecki et al., 2023).

The proposal in Figure 2 is intended merely as a preliminary. First, we will need to revise our classification once we consider the new possibilities raised by machine abilities. Second, there is room for dispute about which categories need to be included and, further, about which organisms belong in which classes. For example, some might argue that sponges or amebae have sensory capacities sufficient to endow them with knowledge and thus that they should be included in the taxonomy (Godfrey-Smith, 2016). Others will disagree.

The value of the approach is that it allows us to explain relevant similarities in behavioral abilities between diverse types of species while at the same time distinguishing between their underlying biological realizers. The same principles will allow us to formulate and explore additional questions of interest, such as what capacities are sufficient for knowledge and which types of animals have these capacities. If we think certain kinds of organisms lack what is needed for the capacity to know, a taxonomy gives us a basis for explanations of their deficiencies in terms of their physical and behavioral similarities along the relevant dimensions.

Our knowledge taxonomy gives us an initial way to understand relationships between different organisms' ways of sensing the world and responding to it, categorizing them differently depending on which properties are relevant. We can highlight the relevance of world models in one kind of context (organisms with brains) into contrast to the importance of certain abilities in another kind of context (flexible behavior in response to sensory input). As always, what categories to include, what belongs in what, and where the lines are drawn admits of degree and debate (do bees have world models?). The beauty of a taxonomy is that it creates a conceptual framework that gives form and focus to such disputes.

Importantly, the classification also demarcates knowers from nonknowers. According to the taxonomy in Figure 2, animals

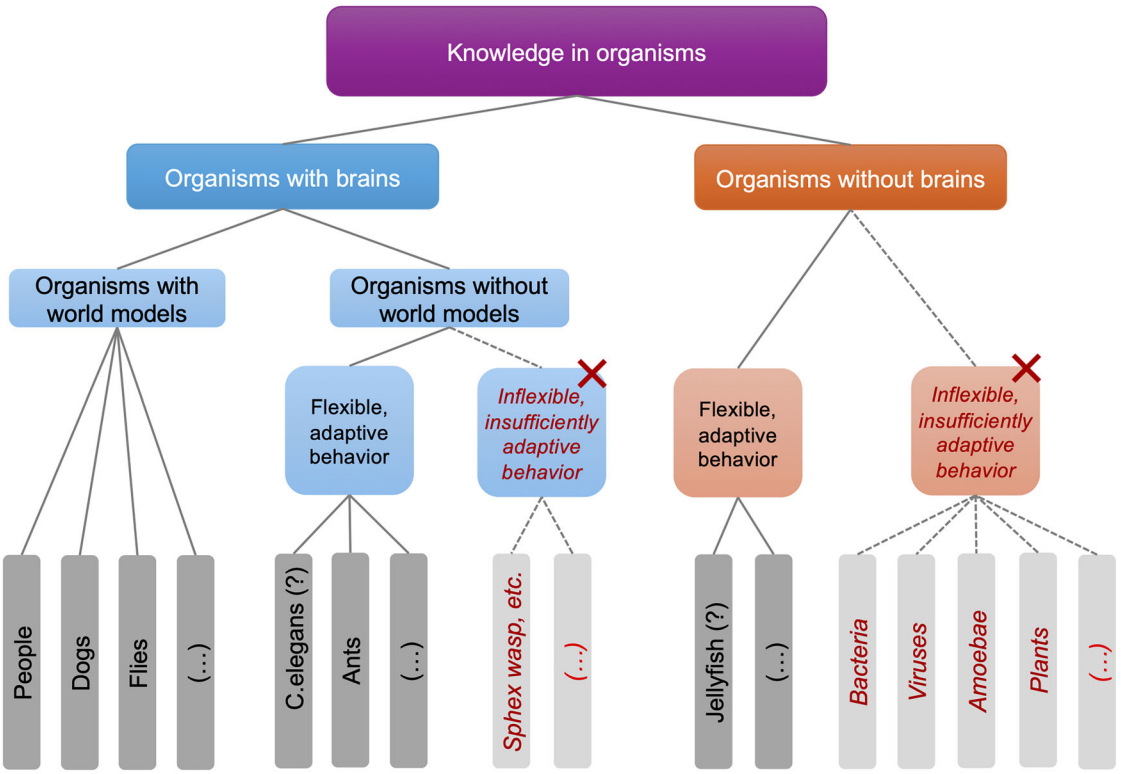

Figure 2. Knowledge in organisms. A preliminary taxonomy of knowledge in organisms with a range of nervous systems. *Items in italics fail to belong to the classification; they are shown as rejected suggestions.*

that lack sufficient abilities to flexibly respond and adapt to their environments lack the capacity for knowledge. It also supports explanations based on this classification; for example, despite having a brain, the sphex wasp cannot know things because it lacks sufficient flexibility in its goal-oriented behavior. By the same token, the taxonomy defines disputes; we might argue that the sphex has been unfairly maligned by showing it can flexibly adapt after all (Keijzer, 2013) or, as I suggested above, argue that the brain of *C. elegans* can support a world model.

The point is not to defend this particular knowledge taxonomy as the right or only one. Perhaps another taxonomy is preferable. In any case, as new neuroscientific facts are discovered and conceptualized, we will need to expand and change this classification. The point is rather that creating a taxonomy allows us to analyze and organize how we think about knowledge. It allows patterns to emerge, identifies gaps, creates debates, and provides structure for further exploration. In particular, to determine whether an organism is capable of knowledge, to explain why it has or why it does not have such capacities, and to predict its possibilities for evolution toward this capacity, we find its place in the taxonomy.

Now we can ask: do LLMs fit into the picture?

LLMs and Knowledge

Our taxonomy of knowledge in Figure 2 explored a classification of knowledge based on how organisms sense and represent the world. It framed knowledge in terms of the capacity to possess and use information to guide actions, implicitly understanding such capacities in terms of neural systems in organisms.

However, our taxonomy has no place for LLM knowledge. LLMs are machines. They lack brains. They do not have neurons (in the biological sense), and they do not sense the world through experience.

So we can't stop here. Rejecting the possibility of knowledge in LLMs simply because they do not fit our original taxonomy seems misguided, a Luddite's response to the advance of science and technology. When ChatGPT tells me that the Koraku-en Garden is in Okayama, correctly answers my complex series of questions about the different types of garden elements it includes, tells me about how to take a bullet train from Tokyo to get there, notes the cultural relevance the garden holds for the Japanese, and even generates new ideas about the garden, it seems right to say that it knows how to correctly answer my questions. We should grant that there is an important sense in which the LLM knows that Koraku-en is a garden in Okayama, even if its way of knowing is different from the way that people and other animals know things.

To develop an account of how it can be correct to say that a chatbot knows things, for example, that a chatbot knows that Koraku-en is a garden in Okayama, we need to try a different approach. We need to explore the possibility of a new taxonomy, augmenting our system in response to the advances of artificial intelligence.

To approach this task, I propose to go back to basics: that is, to start by explicating what the concept of knowledge is. Once we have a conceptual framework in hand, I will explore a way to apply it to machines and propose a taxonomy that can accommodate LLMs. The philosophically traditional way to approach the concept of knowledge develops its relationship to the associated concepts of truth, belief, and justification, so we shall begin there.

Knowledge and Justified True Belief

Traditionally, the philosophical approach to thinking about the concept of knowledge starts with the proposition: "S knows that *p*." In this approach, we start with the basic idea of a person S knowing a proposition *p*. A proposition can be thought of as a sentence

like "The Koraku-en Garden is in Okayama." So an instance of "S knows that *p*" is that Laurie knows that "the Koraku-en Garden is in Okayama." We then break down how knowing this proposition relates to three other concepts: truth, belief, and justification. Developing the relationships between these concepts will help us to understand what knowing a proposition seems to involve. The hope is that performing such an analysis will help us to reason in a more general way about what knowledge requires. [It will not tell us what is sufficient for knowledge, as the analyses have also shown us that we cannot simply reduce knowledge to truth, belief, and justification (Gettier, 1963)].

Start with truth. On the philosophical conception of propositional knowledge, truth is an essential requirement (Gettier, 1963). An agent can only know truths, not falsehoods: we can have false beliefs but not false knowledge. For example, for an agent to know a proposition like "The Koraku-en Garden is in Okayama," it has to be true. I can only know that "The Koraku-en Garden is in Okayama" if it is actually the case that the garden is in Okayama. In our formal framing, we can say *S* knows that *p* only if *p* is true.

Many philosophers have also argued that an agent must believe a proposition in order to know it (Gettier, 1963; Ichikawa and Steup, 2024). Put more intuitively, you can only know what you believe. If I don't believe that "The Koraku-en Garden is in Okayama," how could I know it? If I were asked whether the garden was in Okayama, I would say it wasn't, because I don't believe that it is. All of my behavior and action would support my disbelief: for example, if I wanted to take a train to see the garden, I wouldn't book one to Okayama. Reasoning this way shows that belief in a proposition seems to be needed for an agent to know it. That is, for S to know that *p*, *S* must believe that *p*.

A stronger claim is that S must be justified in believing it. The importance of justification for knowledge is to rule out lucky guesses. Imagine that I randomly point to somewhere on a map of Japan to show you where Koraku-en is, and by sheer luck I point to Okayama and say it is there. Even if what I am saying is true and I believe it, it still doesn't seem that I know that the garden is in Okayama. We can put it this way: for S to know that *p*, S must be justified in believing that *p*.

How do we get the needed justification? The thought is that a belief must be generated in the right way for it to be properly justified. How might a belief be generated in the right way? Many theories have been proposed. Some argue that a belief is justified if it is formed based on evidence and observation. Many add that a belief can be justified only if it is formed by a suitably reliable process (Goldman, 1986). (Lucky guesses are not reliably a way to get things right.) What counts as a reliable process? Roughly, enough of the beliefs produced by this process need to be true. A standard example of a reliable process is perception: a person believes that there is a coffee cup on the table in front of her as the result of a reliable perceptual process, one that was generated in response to actually seeing the coffee cup on the table. Most of the time, for an ordinary person, this process is perfectly reliable, even if the occasional hallucination or illusion could occur.

To summarize, our philosophical analysis tells us that "Laurie knows that the Koraku-en Garden is in Okayama" only if:

1. "The Koraku-en Garden is in Okayama" is true.
2. Laurie believes that "The Koraku-en Garden is in Okayama."
3. Laurie is justified in believing that "The Koraku-en Garden is in Okayama."

More generally, the thought here is that truth, belief, and (reliably formed) justification are deeply related to knowledge, and if a person justifiably believes something true as the result of engaging in a reliable process that forms their belief, we can reasonably infer that they know what they believe. We can reasonably infer that Laurie knows that there is a coffee cup on the table in front of her because (1) there is in fact a coffee cup on the table in front of her, (2) she believes that there is a coffee cup on the table in front of her, and (3) her belief was formed through a reliable perceptual process that was generated in response to the fact that she sees a coffee cup on the table in front of her.

This might seem fine, as far as it goes, but how is such an account relevant to knowledge in LLMs? As far as we know, LLMs are not conscious. As noted above, they don't perceive the world. They don't have physical brains the way that people do or even any type of rudimentary biological neural system. So how could it be meaningful to claim that having beliefs formed under the right conditions is related to how LLMs have knowledge? We are now at the heart of the problem.

Instrumental Knowledge

How can we enrich our philosophical concept to make room for machine knowledge? We must explore new ways to characterize how LLMs could know things, even if they are neither conscious nor biologically grounded. Thus, we need to approach our ordinary philosophical concept of knowledge from a different angle. (This, in turn, will allow us to see how a special kind of belief can also play a role.) While ordinary human knowledge is often grounded by perceptual processes like seeing a coffee cup, there are other ways of knowing things. For example, when you use a TV remote, if you can use it successfully to turn the TV on and off and to select various functions, then you gain a certain kind of knowledge. You know how to turn the TV on, you know how to watch Netflix, and you know how to turn the TV off. In other words, you have instrumental knowledge of how to turn on the TV and watch movies. If you have a thermostat on your wall that you know how to adjust, then you know how to make the house warmer or cooler.

This is a very simple and basic sort of knowing. You don't need to understand how the remote works "under the hood" or anything like that. You don't have to have any idea of how the internal workings of the remote cause the TV to go on or how turning the dial on the thermostat communicates with the heating or cooling system in the house. What you know comes from your successful use of the instrument. That is, you have a kind of "instrumental knowledge" (Yildirim and Paul, 2024a,b).

We can characterize such instrumental knowledge as knowledge acquired by individuals through their ability to reliably and successfully use instruments that perform sufficiently well defined tasks. It is a very minimal, functionally defined, kind of knowledge. It is not very deep. It does not require a capacity for cognition nor for human-level representation. It does not even rise to the level of what many psychologists describe as "instrumental knowledge," as it does not need to be derived from operant conditioning that requires the cognitive capacity to represent the relationship between cause and effect, even if one's behavior is also partly determined by feedback (Dickinson, 1994; Yildirim and Paul, 2024b).

So to have the sort of instrumental knowledge I am defining here, you don't even need to have any deep sense of the larger causal or counterfactual structure involved. For example, you do not need to grasp the counterfactual that, if you had not pressed the "on" button, your remote would not have turned

on the TV. You simply need to demonstrate, through your behavior, that you know what to do in order to watch TV and to modify responses based on reinforcement.

This demonstrative behavior amounts to your implicit grasp of a conditional: you know that "when the 'on' button is pressed, the TV turns on." With your behavior, you demonstrate your implicit grasp of the conditionals needed to use a (reliable) remote well enough to succeed in your task. We can describe such knowledge as highly task-specific: you could not use the remote to do other things, such as to cook a meal or call a friend.

This kind of behaviorally grounded instrumental knowing is obviously quite limited as a species of knowledge, but the very fact of its shallowness means we can measure it rather easily. On this approach, such knowing is not that different from, say, what a small child knows when they first learn how to turn on the TV. As minimal as this type of instrumental knowing is, it gives us what we need to develop a concept of knowledge in machines.

How? We can extend the application of the instrumental knowledge concept by expanding its constituent concept: the concept of an instrument. We don't need to limit the concept of an instrument to physically manipulable things like remote controls or thermostats. We can widen the range of things that we take to be instruments to include more abstractly defined entities.

In particular, language can be treated as an instrument. People use natural language as an instrument when they use it to ask questions and elicit answers. Ordinary speakers use natural language as a tool or an instrument in the following way: "You ask a question of someone [and you] can tell what they think (or say) based on a deliverance conveyed by their utterance. Interpretative knowledge of what a speaker thinks (or says) is thus instrumental knowledge that uses the instrument of language" (Sosa, 2006). Taking language as an instrument opens a new path toward instrumental knowledge.

To be clear, natural language is not the only kind of language that can be used as an instrument, and this is a good thing, because an LLM's ability to generate coherent language responses is not conclusive evidence of understanding natural language. We can say that LLMs can have a kind of formal ability to use language to respond to prompts despite lacking the ability to use natural language (Mahowald et al., 2024). That is, at least at the present stage of scientific development, LLMs lack an ordinary sort of linguistic competence that human beings exhibit. To this end, distinguish between "formal" linguistic competence, defined as knowledge of linguistic rules and patterns, from "functional" linguistic competence, defined as competently understanding and using language in the world. Following Mahowald et al. (2024), "functional" linguistic competence is required for the ability to use natural language. The distinction is grounded on a separation of (mere) linguistic abilities from abstract knowledge and reasoning abilities and supported using evidence from human neuroscience and cognitive science. [For a sampling of the debate over whether or how LLMs could understand and use natural language, see Mitchell and Krakauer (2023), Piantadosi and Hill (2022), Bender et al. (2021), and Mahowald et al. (2024).]

Mahowald et al. (2024) have shown that, unlike humans, LLMs lack functional linguistic competence and, using this conclusion, argue persuasively that LLMs currently lack the capacity for natural language. This, in turn, entails they can't use natural language as an instrument. However, LLMs can be said to have formal linguistic competence, and they seem to have a different tool at their disposal: next-word (or token) generation. That is, we can hold that LLMs use the tool of next-word generation as an instrument (combined with human reinforcement learning and other tools), in order to competently respond with clear and informative answers to prompts.

We can use this to define the way that an LLM could have a kind of instrumental knowledge. It achieves such knowledge to the extent that it can reliably and successfully use the process of next-word generation as its primary instrument. That is, an LLM's wide-ranging and generative ability to reliably and to successfully respond to prompts using next-word generation, to produce correct answers and, even better, productively expand upon these answers, can be treated as demonstrating a kind of instrumental knowledge of the relevant propositions.

[As philosophers will know, this treatment of instrumental knowledge also bears a similarity to discussions in epistemology and the philosophy of language involving expertise and "know-how." I lack the space to develop the relationship between instrumental knowledge and philosophical debates about knowledge-how versus propositional knowledge or "knowledge-that" and so will set this issue aside in this paper. However, my general approach is to embrace the broadly functionalist and inclusive definition of knowledge I gave above. While I am developing an account of LLM knowledge in broadly propositional terms, I am in agreement with the point that "the right response to [those] who are presupposing the intellectualist picture [the view that all knowledge is propositional] is to agree that the paradigm cases of knowing-how do not fit the intellectualist picture of propositional knowledge, but to argue that this is because the traditional intellectualist conception is a wrongheaded account of propositional knowledge, mistaking one narrow range of cases of knowledge-that for the general case" (Stalnaker, 2012, p. 756).]

The conception of instrumental knowledge in LLMs developed in this *Viewpoint* is first introduced in Yildirim and Paul (2024a). The idea, while novel, is a perfectly consistent extension of the original generalization we started with, viz., where we took knowledge to be the possession of information and the capacity to use that information to guide one's actions. To enrich our proposal and defend its place in the knowledge taxonomy, we now return to our earlier exploration of the relationship between knowledge, truth, belief, and justification. How might our conception of instrumental knowledge in LLMs involve these notions?

Truth

Start with the relationship between truth and knowledge. For ordinary propositional knowledge, truth is necessary. ChatGPT can only know that "The Koraku-en Garden is in Okayama" if the Koraku-en garden is actually in Okayama. To bring our LLM concept of instrumental knowledge closer to ordinary propositional knowledge, we should hold that an LLM instrumentally knows that p only if p is true.

That is, ChatGPT can only instrumentally know a proposition if that proposition is true. The need for truth explains why the existence of pervasive chatbot hallucinations undermines their claim to knowledge. If ChatGPT hallucinates and tells me (falsely) that "The Koraku-en Garden is located in Tokyo," it cannot be said to know where Koraku-en is. (It is worth noting that there are other ways of knowing that my account is not addressing directly, e.g., knowing how to write a poem or knowing how to write code. The account of knowledge I am developing can be modified to address these ways of knowing, but I will leave this complication aside in what follows.)

Belief

Let 's turn to belief. Here we find the root of the strangeness of the idea that LLMs can know things. In many ordinary language contexts, to ascribe a belief to an agent requires that the agent has conscious experience or, at least, that the agent represents and understands the world in something like the same way that you and I represent and understand the world.

Here, by hypothesis, we assume chatbots do not have experiences and that they do not represent and understand the world in the way that human beings ordinarily represent and understand the world. We do not even assume they have the capacity for representation in anything more than a strictly computational sense. If chatbots do not represent and understand the world in the way that human beings ordinarily represent and understand the world, and again, by hypothesis, they do not, then how could a chatbot have a belief?

We need to revise our assumptions. The AI revolution has created machines with amazing new abilities. We need to be able to transform our concepts to accommodate this scientific advance (Paul, 2014; Ongchoco et al., 2024). What we have learned from the success of LLMs is that using the tool of next token prediction to exploit patterns drawn from enormous amounts of text can realize a certain kind of ability. This ability, the ability to perform tasks like synthesize arguments from texts, develop conference programs, and give travel information, is complex and generative.

The technique of next-word prediction has shown us that it is possible to build machines that, by hypothesis, do not think and reason the way that human beings can think and reason and yet can reliably generate complex, sophisticated, human-aligned, responsive, and generative answers to prompts. This gives us a whole new way to think about the information that patterns in large amounts of text can carry, and further, to understand what grasping those patterns can mean for the ability to successfully answer questions and perform certain complex tasks.

Before LLMs were created, we'd assumed that having such sophisticated abilities would require an agent to represent and understand the world in much the same way that human beings represent and understand the world. This assumption has been overthrown.

Like the ancient assumption that the earth was the center of the universe, our error was fundamentally anthropocentric. Just as it was mistaken to assume that we were the center of the universe and thus that the sun and planets revolve around the earth, it was anthropocentric to assume that the way humans represent the world using world models is necessary for the ability to summarize complex manuscripts, have conversations, or write poems. We need to reject the implicit assumption that the way a human being represents the world constrains the ways that agents are able to know these complex things. It is anthropocentric to assume that machines need to represent the world the way that human beings do in order for them to know the sophisticated kinds of things that humans can know. (In fact, to the extent that we are willing to say that, e.g., an ant knows its way back to its colony or *C. elegans* knows which temperature gradient to move to, we have already implicitly rejected this anthropocentric assumption.)

Put differently, the abilities of LLMs are part of a scientific revolution, a revolution that brings with it conceptual change. While the AI revolution has brought us many advances and surprises, the discovery that, given enough text and training, a machine could perform these kinds of sophisticated and responsive tasks reveals that the AI revolution is not just a scientific revolution. It is also a conceptual revolution.

The conceptual revolution stems from the surprising discovery that probabilistic techniques used to discover patterns in language (along with human reinforcement learning) can result in machines with sophisticated abilities. The point is that the abilities of these machines are, in essence, "created merely from recognizing patterns in language." LLMs do not have sensory capacities. They do not have brains. The "neural networks" of these machines are not actual networks of neurons, despite the widely used metaphor. They do not seem to have world models (at least, not as of this writing) or employ structure preserving representations of the world even if they can produce text that is human-like. Moreover, they do not learn or use language in the way a person learns and uses a language. At a fundamental level, they are not reasoning or thinking in the way that a human being thinks or reasons, but they can perform as though they were sentient creatures. How could such a thing be possible? And yet, it is.

In this way, the existence of LLMs forces us to reexamine our assumptions about what is necessary for knowledge and for other human-like abilities. Something extraordinary has been created here. In short, we need to transform how we think of knowledge (and related concepts of intelligence) to make conceptual sense of the scientific discovery that these machines embody. We need to do some conceptual engineering (Cappelen, 2018; Ongchoco et al., 2024).

As an example of the role of conceptual engineering in scientific revolution, consider how Einstein, by adopting an interpretation of space based on Minkowski spacetime and the Riemannian theory of manifolds, was able to provide a relativistic treatment of mechanics. This brilliant conceptual move, which required the rejection of the Euclidean interpretation as the universal or default conceptual framework, was the key step toward establishing spacetime as an empirically detectable entity (Minkowski, 1908; Einstein, 1916; Friedman, 2001; Friedman, 2016). For a second example, consider the revolution in organic chemistry around aromatic compounds, which included a newly engineered concept of "resonance" necessary to understand the properties of the benzene molecule (Pauling, 1939; Kekulé, 1866).

For the conceptual engineering needed to accommodate LLM knowledge, the philosophical analysis we performed above, representing decades of philosophical research on the deep connection between knowledge and belief, will serve as a guide. As my analysis suggests, to understand the new way a machine might know things, we need a new account of how a machine can believe things. Recognizing the need for conceptual change that can accommodate the new abilities LLMs demonstrate, we need to refine our conception of belief.

How might we revise our conceptual scheme to include a concept of belief that can pave the way for machine knowledge? We can take beliefs to be states that guide actions that get agents to goals (Stalnaker, 2012). By stipulation, we do not assume that having a belief requires anything as robust as the ability to represent or understand the world in the way that human beings do. As such, the treatment of belief can be expanded to include instrumental beliefs.

Accordingly, define an instrumental belief as a state that guides an individual such that it gains the ability to successfully use an instrument to perform a sufficiently well defined task. Now we can say ChatGPT instrumentally "believes" that the "Koraku-en Garden is in Okayama" in virtue of the fact that it relies on its instrument of next-word prediction to say so in response to the prompt "where is the Koraku-en Garden?"

Justification

Is such a belief justified? Only if it was formed in the right way. To be justified in its belief, we can hold that our LLM must be suitably reliable, i.e., through the reliable success of its highly specific task performance (Goldman, 1979). So we can say that ChatGPT is justified in its instrumental belief that the "Koraku-en Garden is in Okayama" in virtue of the fact that its belief was reliably produced, using its instrument of next-word prediction, in the way it was precisely defined to do so by its engineers.

Reliability is important, and this may be where, in the current scientific context, we want to resist ascribing justification to LLMs like ChatGTP. If we want to say an LLM knows something, it must be able to reliably respond to appropriate prompts with true answers. Such reliability, in certain contexts, continues to elude us. If an LLM cannot reliably generate true answers to appropriate prompts using its instrument of next-word prediction, then then its instrumental beliefs cannot serve as a foundation for its instrumental knowledge.

That said, LLMs are rapidly improving with regard to success on this front, and they are already implemented in a wide variety of academic and corporate contexts where they are successfully performing sets of clearly prescribed tasks (OpenAI, 2023, 2024). Within this circumscribed region, the assumption of reliability is clearly being made, and many would hold that significant improvements will make the instrument of next-word generation sufficiently reliable in the near future.

We can call the justification standard of true answers being reliably generated in a context the standard of "proper use" of the instrument for a context. Even if next-word generation cannot yet be said to meet the standard of proper use for every context, it is already met in many contexts, and strong improvement is expected. Thus, justification is available for many contexts, even if not for all.

With this caveat in hand, our philosophical analysis tells us that "An LLM instrumentally knows that the Koraku-en Garden is in Okayama" only if:

1. "The Koraku-en Garden is in Okayama" is true.
2. The LLM instrumentally believes that "The Koraku-en Garden is in Okayama."
3. The LLM is justified in instrumentally believing that "The Koraku-en Garden is in Okayama."

The upshot here is that an agent with the ability to extract and manipulate patterns in linguistic data in a sophisticated, successful, and generative way realizes a new way to have justified true beliefs, that is, it can have justified, true, instrumental beliefs (for short, *i*beliefs). While, as with ordinary treatments of knowledge, more is needed to get from justified true belief to knowledge, we have opened up a framework for making conceptual sense of how an LLM can know things.

As follows, an LLM can be said to instrumentally believe the answers it produces when they are the product of its instrument of next-word (token) prediction. If these answers meet the justification standard of proper use for reliability and truth, we have conditions that are necessary for the LLM to instrumentally know the answer to the question it was asked. Once we have these conditions in place, we can make better conceptual sense of the claim that a chatbot instrumentally knows that the Koraku-en Garden is in Okayama. For short, it *i*knows that the Koraku-en Garden is in Okayama.

A New Taxonomy: Knowledge in Animals and Machines

Now that we have engineered our conceptual framework, we can propose a new taxonomy. We start by distinguishing kinds of knowledge: in addition to the kind of knowledge that arises from sensory contact with the world, what I shall call "worldly" knowledge, we can have instrumental knowledge, in particular, the type of instrumental knowledge in machines (*i*knowledge) that we have defined above. We need to include both kinds of knowledge in our classification.

We can also use the classification to highlight informative distinctions: if, broadly speaking, machine knowledge is possible, how can an LLM be said to know propositions while a thermostat cannot? The answer: in virtue of the deep differences between their abilities (He et al., 2025). Thermostats, coffee pots, and other simple machines lack the capacity for the sort of adaptive, generative, goal-directed behavior that is necessary for knowledge. We can also use our classification to find relevant similarities to other agents. After all, the toddler only instrumentally knows how to use the remote control, and I have only the slightest instrumental grasp on how to use the copy machine in the office. Dogs know how to press levers. So there is an important similarity between organisms and machines we should capture: each can have *i*knowledge, even if they use different sorts of instruments in different ways to get it.

We now have the means for explaining why, despite making room for the possibility of machine intelligence, a thermostat can only be metaphorically said to "know" what is expected of it when the temperature changes or that a coffee pot "knows" what it is doing when it turns on and so on. Recall how, in our original knowledge taxonomy, we were able to exclude bacteria, plants, and other brainless organisms as they lack the capacity for knowledge due to lacking certain abilities. Using the same principles, we can use our classificatory system to exclude thermostats, coffee pots, and other types of inflexible, insufficiently adaptive machines in a principled way.

We can also distinguish between instrumental knowledge that is the result of operant conditioning and associative learning from the type of instrumental knowledge that we ascribe to LLMs based merely on a grasp of conditionals [for the details of this distinction, interested readers should consult Goddu et al. (2024) and Yildirim and Paul (2024b)]. Last but not least, we can raise important questions within the framework: for example, do jellyfish have worldly knowledge, in virtue of their sensory responses to the world? Or is it rather instrumental knowledge, in virtue of the way they use their neural nets? Or, like people and dogs, could they exhibit both?

Figure 3 captures our new knowledge classifications, including these proposed distinctions and questions.

By expanding our conceptual framework and using it to develop a new taxonomy of knowledge, we have found a place for LLMs. We have thus proposed an answer to the Knowledge Question.

More generally, our taxonomy provides the structure needed to navigate thorny problems arising from the rapid discoveries and developments in machine learning. What kind of behavior is needed for knowledge? Where do we draw the line between knower and non-knower? How much generativity is needed to "guide" behavior that leads to knowledge? How flexible does a machine need to be to qualify as a knower? Does the ability to pass the Turing test suffice (Turing, 1950)? When is a brain necessary? What counts, where do we draw the lines, and which categories

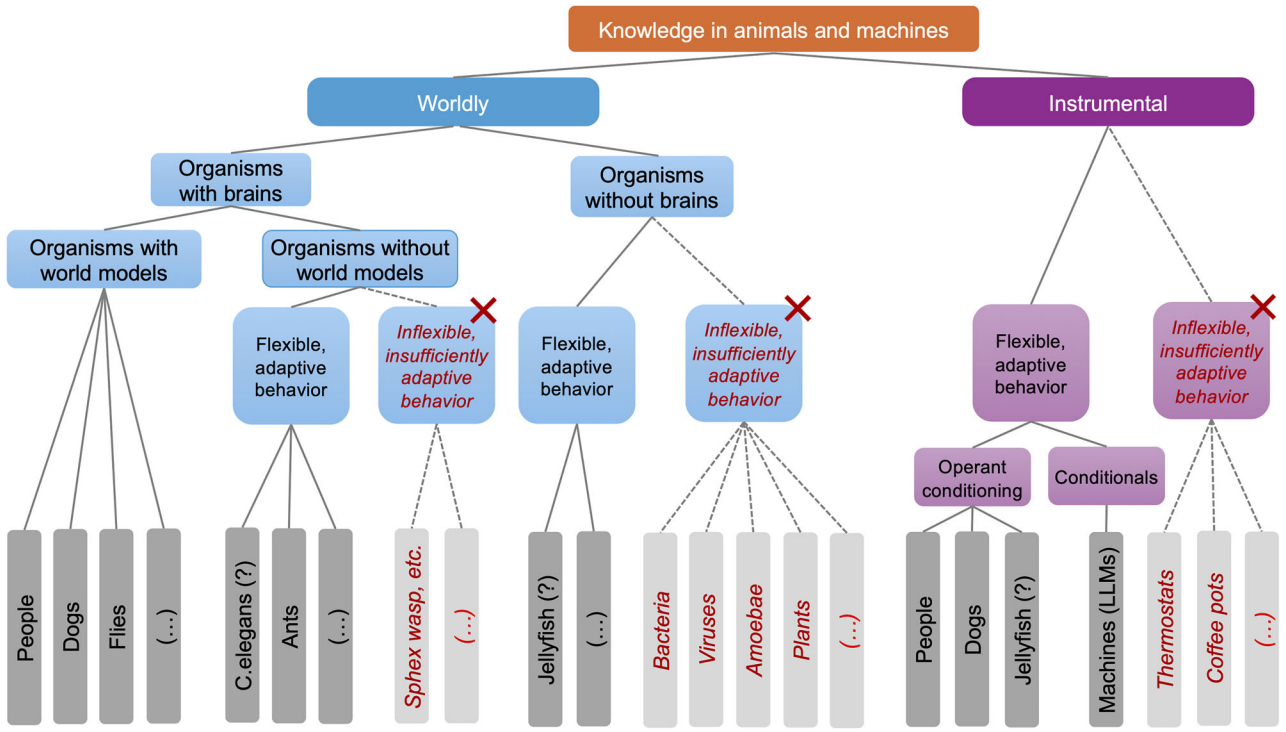

Figure 3. Knowledge in animals and machines. A proposed taxonomy of knowledge that captures different types of knowledge and includes machines like LLMs. *Items in italics fail to belong to the classification; they are shown as rejected suggestions.*

should we include? While debates will continue, and classifications will change as our technology advances, our taxonomy of knowledge provides structure, clarity, and coherence to the discussion.

Human Knowledge Versus Machine Knowledge

Now that we have a framework for knowledge and a proposal for understanding how an LLM can know things, we can use it to generate explanations. In particular, we can use our taxonomy to pinpoint an intuitively felt difference between the way a chatbot knows things and the way that human beings (ordinarily) know things.

Most human knowledge is what I have described as "worldly," based on experience with the world via perception or other types of direct causal engagement. When a person knows where something is located, they know it, at least in part, in virtue of their brain's realizing structure that captures relationships that map or represent the external world, where these relationships are, in turn, generated by the right sort of direct causal contact with the world. You know the coffee cup is on the table through your perceptual experience of it: you see the cup sitting on the table in front of you.

The same thing holds for knowing that the Koraku-en Garden is in Okayama. Even if you have never been to Japan, knowing that the Koraku-en Garden is in Okayama involves knowing something about the world in a very direct sense. You know about gardens as physical things that are located in physical parts of the world (e.g., that a garden is located in Okayama, Japan) at least partly through being embedded in the world as a perceiver. At the very least, you understand the notion of "being located in" through understanding spatial relationships and three-dimensional navigation. You represent the world as the direct result of your physical, perceptual relationship with it, and as a result, the mind's representation of the world matches the world in a way that allows you to know things about it.

Put another way, for a person, knowing that the Koraku-en Garden is in Okayama seems to rely on a particular kind of fit between the mind and the world, generated by the right kind of causal, perceptual engagement with the world. Even if a person learns that the Koraku-en Garden is in Okayama by reading a travel booklet, they know it partly in virtue of their more general capacity to represent the world in a perceptually grounded way. Speaking computationally, a person relies on their world model to know that the Koraku-en Garden is in Okayama.

This is just deeply different from how a chatbot knows that the Koraku-en Garden is in Okayama. An LLM does not have its knowledge in virtue of perceiving the world in any way. It does not perceive things at all. This is a big part of why it can seem strange to say that a chatbot knows that the Koraku-en Garden is in Okayama, even if we are careful to describe such knowledge as instrumental rather than worldly.

Narrowing the Gap

Can we narrow this gap between LLM knowledge and ordinary human knowledge? Is there any way to enrich LLM knowledge so as to bring it closer to the worldly way of knowing?

The best approach to such enrichment, in my view, looks to the science of the mind and the brain. We can explore the theoretical interest of enriching an LLM's representational abilities through incorporating the world models of computational cognitive science (Yildirim and Paul, 2024a). Yildirim and Paul speculate that LLMs could acquire world models with the right causal history and the right structural match to the world, for example, if compression could create representational content. If LLMs could acquire world models, then they would acquire the ability to represent the structure of the world in a way that is directly analogous to how people represent worldly structure. If they could then use their world models to efficiently and reliably generate correct answers to queries, then we can draw stronger and more

interesting parallels between how LLMs know things and how human agents know things. Ultimately, if world models could be incorporated into LLMs, this would develop new capacities for LLMs, bringing them closer to having some sort of richly representational, human-like knowledge.

What such a possibility shows is that worldly knowledge might be achievable through indirect means. That is, by using the representational capacities of world models, we might be able to transmit worldly knowledge to machines, even if such machines lack any direct perceptual contact with the world.

Other Kinds of Knowledge?

Finally, there might be other categories of knowledge that should be included in our taxonomy. For example, people learn things through testimony. Testimonial knowledge in people is passed on through exchanges in natural language, for example, when you learned about the Koraku-en Garden by reading about it from your reliable travel guide. As such, it belongs somewhere in our classification. So we might add testimonial knowledge to our list of kinds of knowing.

One could argue that such testimonial knowledge bears a certain similarity to what LLMs learn by scraping linguistic data from the web. Perhaps, but important caveats apply. Since LLMs lack functional linguistic competence and thus lack the ability to use natural language, they cannot gain testimonial knowledge the way that people gain testimonial knowledge. One could, in response, argue that inferences based on patterns of data could be thought of as gathering a new kind of testimony. That is, while LLMs don't learn from testimony the way that people do, in addition to describing them as having instrumental knowledge, they are capable of in-context learning, and we might also characterize them as learning from testimony in the sense of discovering patterns that emerge from the synthesis of large amounts of text.

Other Debates

My claim that LLMs have the capacity for instrumental knowledge intersects with a related debate (Gopnik, 2022; Yiu et al., 2023), about the abilities of LLMs. Gopnik and her colleagues argue that LLMs are merely "cultural technologies" transmitting or replicating information via text. Others (Lederman and Mahowald, 2024) reject the strong version of this negative claim. They distinguish between "bibliotechnism," which takes LLMs to be mere cultural technologies that cannot realize anything like ordinary human attitudes like beliefs, intentions, or (presumably) knowledge, and a Dennettian "interpretationism," where LLMs, along with a wide variety of agents, can be ascribed attitudes such as beliefs and knowledge simply in virtue of having behavior that is well explained by such ascriptions (Dennett, 1989; Davidson, 2001). Lederman and Mahowald mount a convincing case that LLM abilities to generate novel, meaningful text undercuts bibliotechnism, creating indirect support for a more interpretationist approach.

In other extant work it has been argued, by drawing on the possibility of semantic externalism, that machine utterances could refer in a meaningful way (Cappelen and Dever, 2021; Mandelkern and Linzen, 2024). Part of the story will also involve our ability to attribute algorithms to brains and artificial neural networks (Morrison et al., 2025). Chalmers (2025) delineates the framework and need for propositional interpretability for machines, i.e., for an interpretation of machine mechanisms and behavior in terms of propositional attitudes that human beings can understand. We can also explore the related idea that LLMs could have beliefs, desires, and intentions (Goldstein and Levenstein, 2024). I think all of these authors are mapping out important parts of a rich and nuanced philosophical treatment of LLM attitudes. My account of how LLMs can demonstrate instrumental knowledge of propositions dovetails with these approaches while also stopping short of Dennettian interpretationism.

The problem with full-blown Dennettian interpretationism is that it is too coarse. By design, it individuates attitudes only at the level of behavior and thus cannot generate fine-grained ascriptions of knowledge and corresponding explanations based on differences in underlying physical states that realize the behavior. The deep differences between how human beings perceive and learn about the world and how LLMs interact with the world, reflected in deep differences in the underlying physical states that realize their behaviors, need to be respected in an account of machine knowledge. (Lederman and Mahowald also stop short of full-blown interpretationism. Chalmers makes room for a variety of approaches.)

My contribution to this debate is first, to provide an account of how instrumental knowledge could be possible in LLMs that respects the deep differences between ordinary human knowledge grounded in representation and perception and knowledge grounded merely by the use of instruments, and second, to develop a knowledge taxonomy that clarifies the structure of the debate. My account explains how LLMs can know things in one sense but not in another sense, allows us to distinguish between different kinds of knowledge, and challenges anthropocentric assumptions about belief. It also allows us to draw meaningful parallels between kinds of knowing that humans and machines can share while explaining the divide between them, thus suggesting a new way to think about the classification of knowledge. Moreover, my approach gives us a natural way to bring biologically based neuroscience into the debate about human versus machine intelligence, leveraging classifications based on important underlying physical similarities and differences to make sense of how knowledge is possible across a diverse range of species. Finally, it provides a roadmap for developing additional accounts of related abilities such as memory, learning, and preferring in nonhuman animals and machines.

Discussion

This paper develops a variety of instrumental knowledge, one that is consonant with animal knowledge, yet applicable to machines, the newest entrants to our epistemological kingdom. My approach is intended to support a systematic approach to studying the complexity and diversity of knowledge across a wide spectrum.

Put simply, the analysis develops two different ways we can correctly ascribe knowledge to an agent. Along one dimension of application, what we can call the "instrumental" dimension, to know something requires the ability to reliably, flexibly, and successfully use instruments that implement well defined functions. Along a second dimension of application, what we can call the "worldly" dimension, an agent knows something in virtue of having an underlying physical state caused by interaction with the external environment that represents or corresponds to the physical world in the right way. In *C. elegans*, such physical states are neural states that realize knowing things like which temperature gradient is preferable. In human beings, such physical states include neural states that realize representational ways of knowing in virtue of world models, generated through perception, and exhibiting mind to world "fit." In machines, we can hold that worldly knowledge could be realized by states that incorporate world models.

The idea exploits a multidimensional approach to classification built on the widely accepted practice of distinguishing

different natural kinds at different levels of scientific explanation (Fodor, 1974). A system of natural kinds is a system of classification that reflects the structure of the natural world, grouping things into categories based on what type of thing they are, where types are in turn defined by natural similarities at the intended level of description. Different ways of carving the world up into kinds, corresponding to different levels of scientific description, provide different kinds of explanations.

We used a classification of knowledge kinds to show how animals can be said to know things in worldly ways, ways that machines do not yet exhibit, explaining this in terms of the ways organisms sense and represent the world. At the same time, we've shown that there are important similarities in how animals and machines can both be said to know things, explaining this in terms of similarities in certain abilities. Finally, our approach showed us where to put LLMs into the classification, rigorously defining a sense in which they can know propositions.

By developing our conceptual framework as a knowledge taxonomy, we've gained a better understanding of knowledge itself. In the process, we've uncovered philosophical distinctions that could be of interest to neuroscientists and other experts in biologically based systems of behavior and intelligence. At the same time, we have shown how to preserve the role of the cognitively distinctive nature of human knowledge and the biologically distinctive nature of animal knowledge in our classification of entities with the capacity to know. We have thus articulated a schema for an interdisciplinary discussion about the foundations of knowledge in animals and machines.

References

Bender E, Gebru T, McMillan-Major A, Shmitchell S (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 (FAccT'21), pp 610–623.

Bi W, Lin Q, Peng K, Shah A, Yildirim I (2024) Visual processing of soft objects automatically activates physics-based representations in the human brain. J Vis 24:835.

Bielecki J, Nielsen SK, Nachman G, Garm A (2023) Associative learning in the box jellyfish *Tripedalia cystophora*. Curr Biol 33:4150–4159.e5.

Cappelen H (2018) *Fixing language: an essay on conceptual engineering*. Oxford: Oxford University Press.

Cappelen H, Dever J (2021) *Making AI intelligible: philosophical foundations*. Oxford: Oxford University Press.

Chalmers D (2025) Propositional interpretability in artificial intelligence. https://arxiv.org/abs/2501.15740

Conroy G (2023) Scientists used ChatGPT to generate an entire paper from scratch - but is it any good? Nature 619:443–444.

Davidson D (2001) *Inquiries into truth and interpretation*, Ed 2. Oxford: Clarendon Press.

Dennett D (1989) *The intentional stance*. Cambridge, MA: MIT Press.

Dickinson A (1994) Instrumental conditioning. In: *Handbook of perception and cognition, animal learning and cognition* (Mackintosh NJ, ed), pp 45–79. Cambridge: Academic Press.

Einstein A (1916) Relativity: the special and the general theory. Methuen 1:1–168.

Fodor J (1974) Special sciences (or: the disunity of science as a working hypothesis). Synthese 28:97–115.

Friedman M (2001) *Dynamics of reason: the 1999 kant lectures at Stanford university*. Stanford, CA: CSLI Publications.

Friedman M (2016) *Foundations of space-time theories*. Princeton: Princeton University Press.

Gallistel C (1990) *The organization of learning*. Cambridge, MA: MIT Press.

Gallistel C, King A (2010) *Memory and the computational brain: why cognitive science will transform neuroscience*. Wiley-Blackwell.

Gettier EL (1963) Is justified true belief knowledge? Analysis 23:121–123.

Goddu M, Noë A, Thompson E (2024) LLMs don't know anything: reply to Yildirim and Paul. Trends Cogn Sci 28:963–964.

Godfrey-Smith P (2016) *Other minds: the octopus, the sea, and the deep origins of consciousness*. New York: Farrar, Straus and Giroux.

Goldman A (1986) *Epistemology and cognition*, Vol. 1, pp 1–379. Cambridge: Harvard University Press.

Goldman AI (1979) What is justified belief? In: *Justification and knowledge: new studies in epistemology* (Pappas G, ed.), pp 1–25. Boston: D. Reidel.

Goldstein A, Levenstein J (2024) Does ChatGPT have a mind? https://arxiv.org/abs/2407.11015

Gopnik A (2022) What AI still doesn't know how to do. Wall Street J. https://www.wsj.com/articles/what-ai-still-doesnt-know-how-to-do-11657891316

He R, Cao J, Tan T (2025) Generative artificial intelligence: a historical perspective. Natl Sci Rev 12:nwaf050.

Ichikawa J, Steup M (2024) The analysis of knowledge. In: *The Stanford encyclopedia of philosophy* (Zalta EN, Nodelman U, eds). Palo Alto: Stanford University. https://plato.stanford.edu/archives/fall2024/entries/knowledge-analysis/

Keijzer F (2013) The sphex story: how the cognitive sciences kept repeating an old and questionable anecdote. Philos Psychol 26:502–519.

Kekulé F (1866) On the constitution and metamorphoses of chemical compounds and the chemical nature of carbon. Ann Chem Pharm 137: 129–196.

Khadangi A, Sartipi A, Tchappi I, Fridgen G (2025) CognArtive: large language models for automating art analysis and decoding aesthetic elements. 10.48550/arXiv.2502.04353.

Lederman M, Mahowald K (2024) Are language models more like libraries or like librarians? Bibliotechnism, the novel reference problem, and the attitudes of LLMs.

Luo L, et al. (2014) Dynamic encoding of perception, memory, and movement in a *C. elegans* chemotaxis circuit. Neuron 82:1115–1128.

Luo Y, Li Y, Ogunyemi O, Koski E, Himes BE (2025) Leveraging large language models for academic conference organization. NPJ Digit Med 8: 101.

Mahowald K, Ivanova A, Blank I, Kanwisher N, Tenenbaum J, Fedorenko E (2024) Dissociating language and thought in large language models. Trends Cogn Sci 28:517–540.

Mandelkern M, Linzen T (2024) Do language models' words refer? Comput Linguist 50:1191–1200.

Minkowski H (1908) Space and time. Phys Z 10:75–88.

Mitchell M, Krakauer DC (2023) The debate over understanding in AI's large language models. Proc Natl Acad Sci.U S A 120:e2215907120.

Morrison J, Kriegeskorte N, Peters B (2025) Using transfer learning to define a neural system's algorithm. In Proceedings of the Cognitive Science Society.

Ongchoco DI, Jara-Ettinger J, Paul LA (2024) When new experience leads to new knowledge: a computational framework for formalizing epistemically transformative experiences. Open Mind 8:1291–1311.

OpenAI (2023) GPT-4 technical report. OpenAI Technical Reports, Vol. 4, pp 1–49.

OpenAI (2024) OpenAI o1 system card. https://cdn.openai.com/o1-system-card.pdf

Paul LA (2014) *Transformative experience*. Oxford: Oxford University Press.

Pauling L (1939) *The nature of the chemical bond*, Vol. 1, pp 1–450. Ithaca: Cornell University Press.

Piantadosi S, Hill F (2022) Meaning without reference in large language models. In 36th Conference on Neural Information Processing Systems (NeurIPS).

Schwarz S, Mangan M, Zeil J, Webb B, Wystrach A (2017) How ants use vision when homing backward. Curr Biol 27:401–407.

Sosa E (2006) Knowledge: instrumental and testimonial. In: *The epistemology of testimony* (Lackey J, ed), pp 116–123. Oxford: Oxford University Press.

Stalnaker RC (2012) Intellectualism and the objects of knowledge. Philos Phenomenol Res 85:754–761.

Tolman E (1948) Cognitive maps in rats and men. Psychol Rev 55:189–208.

Turing AM (1950) Computing machinery and intelligence. Mind 59:433–460.

Yildirim I, Paul LA (2024a) From task structures to world models: what do LLMs know? Trends Cogn Sci 28:404–415.

Yildirim I, Paul LA (2024b) Response to Goddu et al.: new ways of characterizing and acquiring knowledge. Trends Cogn Sci 28:965–966.

Yiu E, Kosoy E, Gopnik A (2023) 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.