PNAS PERSPECTIVE

The debate over understanding in AI's large language models

Melanie Mitchell[a,1,2] and David C. Krakauer[a,1]

Edited by Terrence Sejnowski, Salk Institute for Biological Studies, La Jolla, CA; received October 12, 2022; accepted February 14, 2023

We survey a current, heated debate in the artificial intelligence (AI) research community on whether large pretrained language models can be said to understand language—and the physical and social situations language encodes—in any humanlike sense. We describe arguments that have been made for and against such understanding and key questions for the broader sciences of intelligence that have arisen in light of these arguments. We contend that an extended science of intelligence can be developed that will provide insight into distinct modes of understanding, their strengths and limitations, and the challenge of integrating diverse forms of cognition.

artificial intelligence | understanding | large language models

What does it mean to understand something? This question has long engaged philosophers, cognitive scientists, and educators, nearly always with reference to humans and other animals. However, with the recent rise of large-scale AI systems—especially the so-called large language models—a heated debate has arisen in the AI community on whether machines can now be said to understand natural language and thus understand the physical and social situations that language can describe. This debate is not just academic; the extent and manner in which machines understand our world have real stakes for how much we can trust them to drive cars, diagnose diseases, care for the elderly, educate children, and more generally act robustly and transparently in tasks that impact humans. Moreover, the current debate suggests a fascinating divergence in how to think about understanding in intelligent systems, in particular the contrast between mental models that rely on statistical correlations and those that rely on causal mechanisms.

Until quite recently, there was general agreement in the AI research community about machine understanding: While AI systems exhibit seemingly intelligent behavior in many specific tasks, they do not *understand* the data they process in the way humans do. Facial recognition software does not understand that faces are parts of bodies, the role of facial expressions in social interactions, what it means to "face" an unpleasant situation, or any of the other uncountable ways in which humans conceptualize faces. Similarly, speech-to-text and machine translation programs do not understand the language they process, and autonomous driving systems do not understand the meaning of the subtle eye contact or body language drivers and pedestrians use to avoid accidents. Indeed, the oft-noted *brittleness* of these AI systems—their unpredictable errors and lack of robust generalization abilities—are key indicators of their lack of understanding (1). However, over the last several years, a new kind of AI system has soared in popularity and influence in the research community, one that has changed the views of some people about the prospects of machines that understand language. Variously called large language models (LLMs), large pretrained models, or foundation models (2), these systems are deep neural networks with billions to trillions of parameters (weights) that are "pretrained" on enormous natural-language corpora, including large swathes of the web, online book collections, and other collections amounting to terabytes of data. The task of these networks during training is to predict a hidden part of an input sentence—a method called "self-supervised learning." The resulting network is a complex statistical model of how the words and phrases in its training data correlate. Such models can be used to generate natural language, be fine-tuned for specific language tasks (3), or be further trained to better match "user intent" (4). LLMs such as OpenAI's well-known GPT-3 (5) and more recent ChatGPT (6) and Google's PaLM (7) can produce astonishingly humanlike text, conversation, and, in some cases, what seems like human reasoning abilities (8), even though the models were not explicitly trained to reason. How LLMs perform these feats remains mysterious for lay people and scientists alike. The inner workings of these networks are largely opaque; even the researchers building them have limited intuitions about systems of such scale. The neuroscientist Terrence Sejnowski described the emergence of LLMs this way: "A threshold was reached, as if a space alien suddenly appeared that could communicate with us in an eerily human way. Only one thing is clear—LLMs are not human... Some aspects of their behavior appear to be intelligent, but if not human intelligence, what is the nature of their intelligence?" (9).

As impressive as they are, state-of-the-art LLMs remain susceptible to brittleness and unhumanlike errors. However, the observation that such networks improve significantly as their number of parameters and size of training corpora are scaled up (10) has led some in the field to claim that LLMs—perhaps in a multimodal version—will lead to human-level intelligence and understanding, given

Author affiliations: [a]Santa Fe Institute, Santa Fe, NM 87501

Author contributions: M.M. and D.C.K. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]M.M. and D.C.T contributed equally to this work.

[2]To whom correspondence may be addressed. Email: mm@santafe.edu.

Published March 21, 2023.

sufficiently large networks and training datasets. A new AI mantra has emerged: "Scale is all you need" (11, 12).

Such claims are emblematic of one side of the stark debate in the AI research community on how to view LLMs. One faction argues that these networks truly understand language and can perform reasoning in a general way (although "not yet" at the level of humans). For example, Google's LaMDA system, which was pretrained on text and then fine-tuned on dialogue (13), is sufficiently convincing as a conversationalist that it convinced one AI researcher that such systems "in a very real sense understand a wide range of concepts" (14) and are even "making strides toward consciousness" (15). Another machine language expert sees LLMs as a canary in the coal mine of general human-level AI: "There is a sense of optimism that we are starting to see the emergence of knowledge-imbued systems that have a degree of general intelligence" (16). Another group argues that LLMs "likely capture important aspects of meaning, and moreover work in a way that approximates a compelling account of human cognition in which meaning arises from conceptual role" (17). Those who reject such claims are criticized for promoting "AI denialism" (18).

Those on the other side of this debate argue that large pretrained models such as GPT-3 or LaMDA—however fluent their linguistic output—cannot possess understanding because they have no experience or mental models of the world; their training in predicting words in vast collections of text has taught them the *form* of language but not the meaning (19–21). A recent opinion piece put it this way: "A system trained on language alone will never approximate human intelligence, even if trained from now until the heat death of the universe," and "it is clear that these systems are doomed to a shallow understanding that will never approximate the full-bodied thinking we see in humans" (22). Another scholar argued that intelligence, agency, and, by extension, understanding "are the wrong categories" for talking about these systems; instead, LLMs are compressed repositories of human knowledge more akin to libraries or encyclopedias than to intelligent agents (23). For example, humans know what is meant by a "tickle" making us laugh because we have bodies. An LLM could use the word "tickle," but it has obviously never had the sensation. Understanding a tickle is to map a word to a sensation, not to another word.

Those on the "LLMs do not understand" side of the debate argue that while the fluency of large language models is surprising, our surprise reflects our lack of intuition of what statistical correlations can produce at the scales of these models. Anyone who attributes understanding or consciousness to LLMs is a victim of the Eliza effect (24)—named after the 1960s chatbot created by Joseph Weizenbaum that, simple as it was, still fooled people into believing it understood them (25). More generally, the Eliza effect refers to our human tendency to attribute understanding and agency to machines with even the faintest hint of humanlike language or behavior.

A 2022 survey given to active researchers in the natural-language-processing community shows the stark divisions in this debate. One survey item asked whether the respondent agreed with the following statement about whether LLMs could ever, in principle, understand language: "Some generative model [i.e., language model] trained only on text, given enough data and computational resources, could understand natural language in some nontrivial sense." Of 480 people responding, essentially half (51%) agreed, and the other half (49%) disagreed (26).

Those who would grant understanding to current or near-future LLMs base their views on the performance of these models on several measures, including subjective judgment of the quality of the text generated by the model in response to prompts (although such judgments can be vulnerable to the Eliza effect), and more objective performance on benchmark datasets designed to assess language understanding and reasoning. For example, two standard benchmarks for assessing LLMs are the General Language Understanding Evaluation (GLUE) (27) and its successor (SuperGLUE) (28), which include large-scale datasets with tasks such as "textual entailment" (given two sentences, can the meaning of the second be inferred from the first?), "words in context" (does a given word have the same meaning in two different sentences?), and yes/no question answering, among others. OpenAI's GPT-3, with 175 billion parameters, performed surprisingly well on these tasks (5), and Google's PaLM, with 540 billion parameters, performed even better (7), often equaling or surpassing humans on the same tasks.

What do such results say about understanding in LLMs? The very terms used by the researchers who named these benchmark assessments—"general language understanding," "natural-language inference," "reading comprehension," "commonsense reasoning," and so on—reveal an assumption that humanlike understanding is required to perform well on these tasks. But do these tasks actually require such understanding? Not necessarily. As an example, consider one such benchmark, the Argument Reasoning Comprehension Task (29). In each task example, a natural-language "argument" is given along with two statements; the task is to determine which statement is consistent with the argument. Here is a sample item from the dataset:

> **Argument:** Felons should be allowed to vote. A person who stole a car at 17 should not be barred from being a full citizen for life.
>
> **Statement A:** Grand theft auto is a felony.
>
> **Statement B:** Grand theft auto is not a felony.

An LLM called BERT (30) obtained near-human performance on this benchmark (31). It might be concluded that BERT understands natural-language arguments as humans do. However, one research group discovered that the presence of certain words in the statements (e.g., "not") can help predict the correct answer. When researchers altered the dataset to prevent these simple correlations, BERT's performance dropped to essentially random guessing (31). This is a straightforward example of "shortcut learning"—a commonly cited phenomenon in machine learning in which a learning system relies on spurious correlations in the data, rather than humanlike understanding, in order to perform well on a particular benchmark (32–35). Typically, such correlations are not apparent to humans performing the same tasks. While shortcuts have been discovered in several standard benchmarks used to evaluate

language understanding and other AI tasks, many other, as yet undetected, subtle shortcuts likely exist. Pretrained language models at the scale of Google's LaMDA or PaLM models—with hundreds of billions of parameters, trained on text amounting to billions or trillions of words—have an unimaginable ability to encode such correlations. Thus, benchmarks or assessments that would be appropriate for measuring human understanding might not be appropriate for assessing such machines (36–38). It is possible that, at the scale of these LLMs (or of their likely near-future successors), any such assessment will contain complex statistical correlations that enable near-perfect performance without humanlike understanding.

While "humanlike understanding" does not have a rigorous definition, it does not seem to be based on the kind of massive statistical models that today's LLMs learn; instead, it is based on *concepts*—internal mental models of external categories, situations, and events and of one's own internal state and "self". In humans, understanding language (as well as nonlinguistic information) requires having the concepts that language (or other information) describes beyond the statistical properties of linguistic symbols. Indeed, much of the long history of research in cognitive science has been a quest to understand the nature of concepts and how understanding arises from coherent, hierarchical sets of relations among concepts that include underlying causal knowledge (39, 40). These models enable people to abstract their knowledge and experiences in order to make robust predictions, generalizations, and analogies; to reason compositionally and counterfactually; to actively intervene on the world in order to test hypotheses; and to explain one's understanding to others (41–47). Indeed, these are precisely the abilities lacking in current AI systems, including state-of-the-art LLMs, although ever-larger LLMs have exhibited limited sparks of these general abilities. It has been argued that understanding of this kind may enable abilities not possible for purely statistical models (48–52). While LLMs exhibit extraordinary *formal linguistic competence*—the ability to generate grammatically fluent, humanlike language—they still lack the conceptual understanding needed for humanlike *functional* language abilities—the ability to robustly understand and use language in the real world (53). An interesting parallel can be made between this kind of functional understanding and the success of formal mathematical techniques applied in physical theories (54). For example, a long-standing criticism of quantum mechanics is that it provides an effective means of calculation without providing conceptual understanding.

The detailed nature of human concepts has been the subject of active debate for many years. Researchers disagree on the extent to which concepts are domain-specific and innate versus more general-purpose and learned (55–60), the degree to which concepts are grounded via embodied metaphors (61–63) and are represented in the brain via dynamic, situation-based simulations (64), and the conditions under which concepts are underpinned by language (65–67), by social learning (68–70), and by culture (71–73). In spite of these ongoing debates, concepts, in the form of causal mental models as described above, have long been considered to be the units of understanding in human cognition. Indeed, the trajectory of human understanding—both individual and collective—is the development of highly compressed, causally based models of the world analogous to the progression from Ptolemy's epicycles to Kepler's elliptical orbits and to Newton's concise and causal account of planetary motion in terms of gravity. Humans, unlike machines, seem to have a strong innate drive for this form of understanding both in science and in everyday life (74). We might characterize this form of understanding as requiring few data, minimal or parsimonious models, clear causal dependencies, and strong mechanistic intuition.

The key questions of the debate about understanding in LLMs are the following: 1) Is talking of understanding in such systems simply a category error, mistaking associations between language tokens for associations between tokens and physical, social, or mental experience? In short, is it the case that these models are not, and will never be, the kind of things that can understand? Or conversely, 2) do these systems (or will their near-term successors) actually, even in the absence of physical experience, create something like the rich concept-based mental models that are central to human understanding, and, if so, does scaling these models create ever better concepts? Or, 3) if these systems do not create such concepts, can their unimaginably large systems of statistical correlations produce abilities that are functionally equivalent to human understanding? Or, indeed, that enable new forms of higher-order logic that humans are incapable of accessing? And at this point will it still make sense to call such correlations "spurious" or the resulting solutions "shortcuts?" And would it make sense to see the systems' behavior not as "competence without comprehension" but as a new, nonhuman form of understanding? These questions are no longer in the realm of abstract philosophical discussions but touch on very real concerns about the capabilities, robustness, safety, and ethics of AI systems that increasingly play roles in humans' everyday lives.

While adherents on both sides of the "LLM understanding" debate have strong intuitions supporting their views, the cognitive science–based methods currently available for gaining insight into understanding are inadequate for answering such questions about LLMs. Indeed, several researchers have applied psychological tests—originally designed to assess human understanding and reasoning mechanisms—to LLMs, finding that LLMs do, in some cases, exhibit humanlike responses on theory-of-mind tests (14, 75) and humanlike abilities and biases on reasoning assessments (76–78). While such tests are thought to be reliable proxies for assessing more general abilities in humans, they may not be so for AI systems. As we described above, LLMs have an unimaginable capacity to learn correlations among tokens in their training data and inputs, and can use such correlations to solve problems for which humans, in contrast, seem to apply compressed concepts that reflect their real-world experiences. When applying tests designed for humans to LLMs, interpreting the results can rely on assumptions about human cognition that may not be true at all for these models. To make progress, scientists will need

to develop new kinds of benchmarks and probing methods that can yield insight into the mechanisms of diverse types of intelligence and understanding, including the novel forms of "exotic, mind-like entities" (79) we have created, perhaps along the lines of some promising initial efforts (80, 81).

The debate over understanding in LLMs, as ever larger and seemingly more capable systems are developed, underscores the need for extending our sciences of intelligence in order to make sense of broader conceptions of understanding for both humans and machines. As neuroscientist Terrence Sejnowski points out, "The diverging opinions of experts on the intelligence of LLMs suggests that our old ideas based on natural intelligence are inadequate" (9). If LLMs and related models succeed by exploiting statistical correlations at a heretofore unthinkable scale, perhaps this could be considered a novel form of "understanding", one that enables extraordinary, superhuman predictive ability, such as in the case of the AlphaZero and AlphaFold systems from DeepMind (82, 83), which respectively seem to bring an "alien" form of intuition to the domains of chess playing and protein structure prediction (84, 85).

It could thus be argued that in recent years, the field of AI has created machines with new modes of understanding, most likely new species in a larger zoo of related concepts, that will continue to be enriched as we make progress in our pursuit of the elusive nature of intelligence. And just as different species are better adapted to different environments, our intelligent systems will be better adapted to different problems. Problems that require enormous quantities of historically encoded knowledge where performance is at a premium will continue to favor large-scale statistical models like LLMs, and those for which we have limited knowledge and strong causal mechanisms will favor human intelligence. The challenge for the future is to develop new scientific methods that can reveal the detailed mechanisms of understanding in distinct forms of intelligence, discern their strengths and limitations, and learn how to integrate such truly diverse modes of cognition.

ACKNOWLEDGMENTS. This material is based in part upon work supported by the Templeton World Charity Foundation and by the National Science Foundation under grant no. 2020103. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.

1. M. Mitchell, Artificial intelligence hits the barrier of meaning. *Information* **10**, 51 (2019).
2. R. Bommasani *et al*., On the opportunities and risks of foundation models. arXiv [Preprint] (2021). http://arxiv.org/abs/2108.07258 (Accessed 7 March 2023).
3. B. Min *et al*., Recent advances in natural language processing via large pre-trained language models: A survey. arXiv [Preprint] (2021). http://arxiv.org/abs/2111.01243 (Accessed 7 March 2023).
4. L. Ouyang *et al*., Training language models to follow instructions with human feedback. arXiv [Preprint] (2022). http://arxiv.org/abs/2203.02155 (Accessed 7 March 2023).
5. T. Brown *et al*., Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
6. J. Schulman *et al*., ChatGPT: Optimizing language models for dialogue. *UpToDate* (2022). https://openai.com/blog/chatgpt. Accessed 7 March 2023.
7. A. Chowdhery *et al*., PaLM: Scaling language modeling with Pathways. arXiv [Preprint] (2022). http://arxiv.org/abs/2204.02311 (Accessed 7 March 2023).
8. J. Wei *et al*., Chain of thought prompting elicits reasoning in large language models (2022). http://arxiv.org/abs/2201.11903 (Accessed 7 March 2023).
9. T. Sejnowski, Large language models and the reverse Turing test. arXiv [Preprint] (2022). http://arxiv.org/abs/2207.14382 (Accessed 7 March 2023).
10. J. Wei *et al*., Emergent abilities of large language models. arXiv [Preprint] (2022). http://arxiv.org/abs/2206.07682 (Accessed 7 March 2023).
11. N. de Freitas, 14 May 2022. https://twitter.com/NandoDF/status/1525397036325019649. Accessed 7 March 2023.
12. A. Dimakis, 16 May 2022. https://twitter.com/AlexGDimakis/status/1526388274348150784. Accessed 7 March 2023.
13. R. Thoppilan *et al*., LaMDA: Language models for dialog applications. arXiv [Preprint] (2022). http://arxiv.org/abs/2201.08239 (Accessed 7 March 2023).
14. B. A. y Arcas, Do large language models understand us? *UpToDate* (2021). http://tinyurl.com/38t23n73. Accessed 7 March 2023.
15. B. A. y Arcas, Artificial neural networks are making strides towards consciousness. *UpToDate* (2022). http://tinyurl.com/ymhk37uu. Accessed 7 March 2023.
16. C. D. Manning, Human language understanding and reasoning. *Daedalus* **151**, 127–138 (2022).
17. S. T. Piantasodi, F. Hill, Meaning without reference in large language models. arXiv [Preprint] (2022). http://arxiv.org/abs/2208.02957 (Accessed 7 March 2023).
18. B. A. y Arcas, Can machines learn how to behave? *UpToDate* (2022). http://tinyurl.com/mr4cb3dw (Accessed 7 March 2023).
19. E. M. Bender, A. Koller, 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* (2020), pp. 5185–5198.
20. E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big? in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), pp. 610–623.
21. G. Marcus, Nonsense on stilts. Substack, 12 June 2022. https://garymarcus.substack.com/p/nonsense-on-stilts.
22. J. Browning, Y. LeCun, AI and the limits of language. *UpToDate* (2022) https://www.noemamag.com/ai-and-the-limits-of-language. Accessed 7 March 2023.
23. A. Gopnik, What AI still doesn't know how to do. *UpToDate* (2022). https://www.wsj.com/articles/what-ai-still-doesnt-know-how-to-do-11657891316. Accessed 7 March 2023.
24. D. R. Hofstadter, *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought* (Basic Books, Inc., New York, NY, 1995).
25. J. Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (WH Freeman & Co, 1976).
26. J. Michael *et al*., What do NLP researchers believe? Results of the NLP community metasurvey. arXiv [Preprint] (2022). http://arxiv.org/abs/2208.12852 (Accessed 7 March 2023).
27. A. Wang *et al*., "GLUE: A multi-task benchmark and analysis platform for natural language understanding" in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (Association for Computational Linguistics, 2018), pp. 353–355.
28. A. Wang *et al*., SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Adv. Neural Inf. Process. Syst.* **32**, 3266–3280 (2019).
29. I. Habernal, H. Wachsmuth, I. Gurevych, B. Stein, "The argument reasoning comprehension task: Identification and reconstruction of implicit warrants" in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2018), pp. 1930–1940.
30. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding" in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2019), pp. 4171–4186.
31. T. Niven, H.-Y. Kao, Probing neural network comprehension of natural language arguments" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), pp. 4658–4664.
32. R. Geirhos *et al*., Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
33. S. Gururangan *et al*., "Annotation artifacts in natural language inference data" in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2018), pp. 107–112.
34. S. Lapuschkin *et al*., Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.* **10**, 1–8 (2019).
35. R T. McCoy, E. Pavlick, T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), pp. 3428–3448.
36. S. R. Choudhury, A. Rogers, I. Augenstein, Machine reading, fast and slow: When do models 'understand' language? arXiv [Preprint] (2022). http://arxiv.org/abs/2209.07430 (Accessed 7 March 2023).
37. M. Gardner *et al*., "Competency problems: On finding and removing artifacts in language data" in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (2021).
38. T. Linzen, How can we accelerate progress towards human-like linguistic generalization? in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), pp. 5210–5217.
39. C. Baumberger, C. Beisbart, G. Brun, "What is understanding? An overview of recent debates in epistemology and philosophy of science" in *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science* (Routledge, 2017), pp. 1–34.
40. J. L. Kvanvig, "Knowledge, understanding, and reasons for belief" in *The Oxford Handbook of Reasons and Normativity* (Oxford University Press, 2018), pp. 685–705.
41. M. B. Goldwater, D. Gentner, On the acquisition of abstract knowledge: Structural alignment and explication in learning causal system categories. *Cognition* **137**, 137–153 (2015).