Language Models Understand Us, Poorly

Jared Moore
University of Washington School of Computer Science
jared@jaredmoore.org

Abstract

Some claim language models understand us. Others won't hear it. To clarify, I investigate three views of human language *understanding*: *as-mapping*, *as-reliability* and *as-representation* (§2). I argue that while behavioral reliability is necessary for understanding, internal representations are sufficient; they climb the right hill (§3). I review state-of-the-art language and multi-modal models: they are pragmatically challenged by underspecification of form (§4). I question the *Scaling Paradigm*: limits on resources may prohibit scaled-up models from approaching understanding (§5). Last, I describe how *as-representation* advances a science of understanding. We need work which probes model internals, adds more of human language, and measures what models can learn (§6).

1 Introduction

A theme of EMNLP this year is "unresolved issues in NLP." Hence I consider what it means to understand human language, whether current language models understand and whether future models will.

Recent large language models have achieved impressive results on benchmark tasks (Thoppilan et al., 2022; Brown et al., 2020). These results challenge ordained wisdom on the representations necessary for language production. We've seen improved results from multi-modal models (Saharia et al., 2022; Ramesh et al., 2022, 2021; Shuster et al., 2020; Radford et al., 2022; Borsos et al., 2022), what some call foundation models (Bommasani et al., 2021). Some models even run images, text, and games (Reed et al., 2022). Michael et al. (2022) identify language understanding and scaling as pertinent and much debated questions in NLP.

So what's next? I identify three views on language understanding (§2): *understanding-as-mapping*, *understanding-as-reliability*, and *understanding-as-representation*. Through examples of recent limitations of language models (§4), I argue for *understanding-as-representation* because it climbs the right hill (§3). In particular, I question the assumption that scaling current models is computationally feasible to lead to human-like understanding (§5). Because of the large gap between human and model understanding, I think it is generally misapplied to say that models "understand" (§6.1). Better applied are examples of promising work on understanding (§6.2).

2 Views on Understanding

Some argue that there is a strict barrier which separates human from machine understanding (Bender and Koller, 2020; Searle, 1980). **Understanding-as-mapping** puts *understanding in terms of an absolute mapping between form and meaning*. Here, meaning comes from what a series of forms describes. Those forms can be composed in a variety of ways to yield different, legible meanings.[1] Often, those with this view imply humans have special access to meaning.

Others argue that we ought be rid of the distinction between human and machine understanding. They imply models will close the gap soon enough (Manning, 2022; Agüera y Arcas, 2022; Kurzweil, 2005; Turing, 1950). **Understanding-as-reliability** puts *understanding as a question of reliable communication*: can one agent expect another agent to respond to stimuli in a certain way?[2] This view assumes that scaling alone will lead to an agent capable of human-like language; system internals don't matter. For example, in the most extreme case we can imagine a very large look-up table with state (cf. Russell and Norvig 2021): a mapping from every input sequence to a sensible output sequence.

In this paper, I put *understanding in terms of internal, dynamical representation*: when

[1] Goldberg (2015) reviews compositionality.
[2] Michael (2020) names this the behaviorist view.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 214–222
December 7-11, 2022 ©2022 Association for Computational Linguistics

prompted with a stimulus, does an agent reproduce an internal representation similar enough to that intended? Call this **understanding-as-representation**. Many have proposed related theories (Shanahan and Mitchell, 2022; Barsalou, 2008; Hofstadter and Sander, 2013; Jackendoff et al., 2012; Grice, 1989). In this view, if someone unthinkingly blurts out the correct answer to a question, they would not have understood. While a thermostat reproduces a certain representation given a temperature this representation is not similar to a person's. Some have said that models appear not to understand because their interrogators fail to present stimuli in a model-understandable way (Michael 2020 summarizes). Exactly: I am concerned with human language understanding–not any possible form of understanding.

To advance a science of understanding, I argue that *as-reliability* is necessary, *as-representation* is sufficient, and *as-mapping* is neither.

I reject the premise of *as-mapping* that the way we use words is separate from our meanings. While current work in NLP poorly approximates shared intentionality[3] I disagree that this is the only route to meaning.[4] We could imagine a very large look-up table. There is no boundary between what is and what is not a language.[5]

I accept *as-reliability* in theory. Enough data and parameters should yield a language-performant agent indistinguishably similar to a human tested on byte streams passed along a wire. Similarly, Potts (2022) argues that a self-supervised foundation model could do so. Still, I am skeptical of what I call the *Scaling Paradigm*, that scale alone is a realistic approach.

I think that hill climbing works but we're climbing the wrong hill.

3 Climbing the Right Hill

As-representation and *as-reliability* are compatible: we may care about *representation* but more easily look for *reliability*. I argue that input-output behavioral tests are necessary but may not be sufficient to attribute understanding–we may need to look inside.[6]

[3]The meaning to which Bender and Koller (2020) says models have no access.

[4]Millikan (2017) offers an account where inner representations exist but are not shared.

[5]Bender and Koller (2020) permit meaning in models which ground linguistic form on images.

[6]Compare Churchland and Churchland (1990).

Nonetheless, Alisha, when messaging with Bowen, has no need to look inside Bowen's head to verify that he understood the following exchange:

A: I'm unhappy.
B: Why aren't you happy?

Our human bias is to assume that other agents understand until evidence proves otherwise (Weizenbaum, 1976). This is pragmatic; until recently humans did not encounter non-human agents who could respond somewhat reliably. *Humans assume a similarity of representation*, that others have the same inductive biases.

We can't make that assumption with our models. We can't assume that a chat-bot has a bias to coo over babies (cf. Hrdy 2009). This is why Turing's (1948) test doesn't work–the smoke and mirror programs which won the Loebner prize unintentionally parody input-output tests (Minsky, 1995). Reliability, while useful, alone does not advance a science of understanding. *As-reliability* does not tell us which biases induce understanding. It is not causal.

Granted, humans' internal representations are difficult to measure, may change at each point of access, and in AI we've historically leaned too heavily on certain putative representations. Sutton (2019) calls this a "bitter lesson."

So why talk of representation? I agree with the "bitter lesson" but I also know that there is no such thing as free lunch; human language occupies a small manifold in the space of possible functions. I don't argue to replicate natural functions but rather to be honest about human strengths lest we wander off into fruitless regions of state space. To do logic, at some internal level a system is going to have to appear to use the parts of logic.

Advancing *as-representation* does not mean we know what representations underlie human language nor that we must use certain ones.

Advancing *as-representation* does mean that we pay attention to the constraints on human language usage (§4). We should use those to guide our benchmark tests for reliability. We should not get lost in our proxies, especially what the *Scaling Paradigm* assumes (§5).

4 Under-specification of Meaning

Language is dynamic (e.g. has a history), intersubjective (multi-agent), grounded in a large number of modalities (senses), collectively intentional (in a cultural context), and more. Present models have little, if any, data on these aspects of language.

As Bisk et al. (2020) make clear in their "world scopes," the majority of work in NLP attempts to learn language from internet text alone. I agree that models have fewer data of the world than humans (Bender and Koller, 2020). What our models see under-specify our meanings.

Recent work has looked into such dissimilarities. McCoy et al. (2021) note that while language models use novel constructions, they copy from training data at a high rate. McCoy et al. (2019) and Branco et al. (2021) identify how models use heuristics and short-cuts contrary to the human meaning of prompts. Shaham et al. (2022) show that a current model will not remember a game after a long-enough context window.

Work has just begun to show limitations of multi-modal models. Thrush et al. (2022) test for compositionality over images and text: current models perform at chance. Marcus et al. (2022) and Conwell and Ullman (2022) show many similar compositional issues for DALL-E 2 in comparison to humans. Lake and Murphy (2021) note the implausibility of present multi-modal models: e.g. they cannot describe internal desires or change beliefs.

Thus my claim is not that models can't learn meaning. My claim is that for models to approach human meaning they will require data on aspects of language the field has only begun to investigate. Consider two examples of under-specification:

Under-specification of physics. A model over text and static images would perform poorly on a query such as, *"Can you remove this block without causing the tower to fall?"* paired with an image where a finger points at a block that could obviously be removed (or obviously not).

Under-specification of time. In Western contexts respondents associate earlier to the left and later to the right. This is not specified in language and is mutable (Casasanto and Bottini 2014, and for the example). Thus I expect models which have no notion of time to perform poorly on tests of temporal bias such as those in Fig. 1.

5 Challenges to the *Scaling Paradigm*

But what of more data? The *Scaling Paradigm* tells us that scale alone–more parameters, more data, more modalities–will be enough to approximate human language understanding. Exponential increases in parameters or training data have yielded

Prompt: are each of these events temporally earlier (A), present (B), or later (C)?

one day after			*a day before*		
A	B	**(C)**	**(A)**	B	C
retfa yad eno			*erofeb yad a*		
(C)	B	A	C	B	**(A)**

Figure 1: Stimuli to probe temporal biases. Answers **(bolded)**. In human subjects response times are shorter, once trained, for pairings shown and longer when the order of the answers is reversed. A model with similar temporal bias should assign a higher probability to the correct answer in the intuitive ordering (shown).

linear increases in performance (Chowdhery et al., 2022; Kaplan et al., 2020).

Still, text doesn't seem like enough. Merrill et al. (2021a) argue that there aren't enough examples to learn meaning from form in languages of assertions. Even Chowdhery et al. (2022, pg. 48) admit to be running out of clean data for exponentially-bigger models. Furthermore, by the age of five, the average American child has heard between ten and fifty million words (Sperry et al., 2019). A state-of-the-art model sees from 10k to 100k more words than a kid.[7] For some (e.g. Linzen 2020), the argument stops there: our language models must not be learning the correct functions because they require more data to generalize.

Nonetheless, humans have plenty of data to compare language use besides the words they hear (Tomasello, 2003; Lakoff and Johnson, 1980). For example, Smith et al. (2018) show that children at first need items to be visually centered to learn them. So the claim that models learn the wrong function may only apply when limited to text.

What of more modalities, then? To extrapolate on the figures of PaLM (Chowdhery et al., 2022, §13), we only need to scale a model to about 2^{12} billion parameters and train it for 2.59×10^{26} flops (and increase training data) for perfect performance on a variety of English NLP tasks. For perspective, Heim (2022) estimates the cost of PaLM at 10 million USD which, at the 100x projected, is 1 billion. These figures seem infeasible.

Furthermore, the tests in PaLM do not capture the limits I mentioned for human-like understanding (§4); we would need to add image (Ramesh et al., 2022) and game (Reed et al., 2022) networks

[7] GPT-3 uses .3 trillion tokens (Brown et al., 2020), LaMDA 2.8 trillion (Thoppilan et al., 2022), and PaLM .7 (Chowdhery et al., 2022).

as well. I don't know exactly what effect these added modalities will have except that they will increase the exponent on scale. I am wary that the long tail of returns embraced by the *Scaling Paradigm* chases an exponential.[8] The primacy of scale implies an exponentially diminishing longtail of capability. How soon until our models become planetary in size, as Bostrom (2014) portends?

I thus recast the debate, not as what can theoretically be learned from data (as the *Scaling Paradigm* trumpets), but as the computational efficiency of different learning approaches. *To asymptotically approach an approximation of human language understanding, how many parameters, data, and modalities will our models need?*

These concerns about scale make me hesitant to suggest that efforts will soon close the gap between human and machine understanding (contra *as-reliability*) even as I agree that they will narrow it (contra *as-mapping*).

6 Discussion

6.1 Sorta Understands != Understands

To admit models may one day understand like us and to claim they now understand in limited domains, need not lead us to give understanding over to machines far to one end of the spectrum. These models understand parts of language—just not in the same way as humans understand each other. They are not biased in the same way humans are. Machines will make unknown mistakes unless we can interrogate their representations.

Moreover, attributing "understanding" is consequential. Some argue for granting moral rights to robots (Gordon and Gunkel, 2021). Page describes those decrying AI as "speciesist" (Tegmark, 2017). Bryson (2010) argues to the contrary. But note that understanding and moral status aren't the same. I urge caution over assuming the distinction between human and machine will disappear anytime soon.

We should not let the theoretical capacities of AI blind us to present realities. Saying that current large language models understand is, as McDermott (1976) described a while back, another case of a "wishful mnemonic."

[8]While some note the exponential scales of large models (Thompson et al., 2021; John and Musser, 2022) they may not account for counter-measures (Patterson et al., 2022).

6.2 Pragmatic NLP

Instead of denying or abandoning understanding, *representation* advances a science. It allows us to answer: how does this system understand? How similar are the representations of these two systems? This is what Harnad (1990) and Santoro et al. (2022) call for. Indeed, *as-representation* describes emerging trends in NLP which ...

Probe model internals. While most benchmark tasks focus on input-output reliability (Linzen, 2020; Zhang et al., 2021), investigating understanding will require functional analysis. Buckner (2020) calls for us to determine a taxonomy of the non-robust features detected by deep nets. Beckers et al. (2020) show how to compare causal models at multiple levels of granularity. Geiger et al. (2021), Li et al. (2021), and Lovering and Pavlick (2022) extend this analysis with interventions to ask which representations (simpler models) approximate large language models. Olsson et al. (2022) find evidence for so-called induction heads in transformer models. Johnston and Fusi (2022) find abstract "representations" emergent from neural networks trained on similar tasks. Merrill et al. (2021b) investigate norm growth saturation in transformers as their inductive biases.

Add more of human language such as intersubjective, multi-agent environments. Noukhovitch et al. (2021) find that partially-competitive agents learn to use symbols. In a text game, Hendrycks et al. (2021) gauge moral valence. This is as Firestone (2020) describes, to use species-fair human-machine comparisons.

We might focus on human data constraints, such as the CHILDES database of language learning (MacWhinney, 2000; Linzen, 2020). Hill et al. (2020) show how the increased modality of data to a deep network may lead to generalizability. Contra strict composition, Santoro et al. (2022) argue for the inclusion of "socio-cultural interactions." These are similar to calls for dynamic grounding (Chandu et al., 2021) and "common sense" (Sap et al., 2020). For example,we need models which not only resolve gaze (Koleva et al., 2015) but also deploy sharing gaze to sharing in other domains.

On generalizability Mitchell (2019) asks us to consider micro domains: "abc:abd; xxyyzz:?". Şahin et al. (2020) propose a task with a small dataset from Rosetta Stone. Brachman and Levesque (2022) propose using only the kids version of Wikipedia, KidzSearch, to reduce the number of entities on which to train. A language performant agent should be able to do well in this micro-domain alone.

Measure what models can learn. We need more work on the tractability of meaning along the lines of Merrill et al. (2021a) on a language of assertions. How many different streams of data (or "world scopes" Bisk et al. 2020) must we add to models to make them more reliable? What kind of scaling factors can we reasonably expect as we include more aspects of human language?

7 Conclusion

Present language models have limited access to meaning (§4) and scale alone may not be sufficient to achieve human-level understanding (§5), at least until we can guarantee similar representations or inductive biases (§2).

Inference tasks reduce the space of possible entities implied by conventional usage. Still, we must be assured of what they represent in order to guarantee their reliability. *Understanding-as-mapping* would deny the distributional compositionality of our languages and of our minds. *Understanding-as-reliability* would claim our machines "get" meaning so long as we focus only on temporally limited usage. *Understanding-as-representation* would focus on accurate, measurable, and tractable meaning.

Limitations

This is only one perspective of many possible on the nature of understanding. Given the short form of this presentation I am not able to do justice to the diverse fields–from philosophy of science to linguistics to brain science–which I reference. I may therefore insufficiently explain the relevant literature to my target readers in NLP. While I have attempted to present a well-informed prior on the future direction of AI, my perspective is nonetheless uncertain; I may ultimately be incorrect.

Acknowledgements

Julian Michael, Ari Holtzman, Dallas Card, Johan Michalove, and Blaise Agüera y Arcas provided the invaluable conversations, questions, and challenges that have made the ideas of this paper what they are. The participants of Melanie Mitchell's "Embodied, Situated, and Grounded Intelligence" workshop at the Santa Fe Institute helped me realize the need for work such as this. I would also like to thank suggestions of my three anonymous reviewers.

References

Blaise Agüera y Arcas. 2022. Do large language models understand us? *Dædalus*, 151(2).

Lawrence W. Barsalou. 2008. Grounded Cognition. *Annual Review of Psychology*, 59(1):617–645.

Sander Beckers, Frederick Eberhardt, and Joseph Y. Halpern. 2020. Approximate Causal Abstractions. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, pages 606–615. PMLR. ISSN: 2640-3498.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience Grounds Language. *arXiv:2004.10151 [cs]*. ArXiv: 2004.10151.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258 [cs]*. ArXiv: 2108.07258.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2022. AudioLM: a Language Modeling Approach to Audio Generation. ArXiv:2209.03143 [cs, eess].

Nick Bostrom. 2014. *Superintelligence: paths, dangers, strategies*, first edition edition. Oxford University Press, Oxford.

Ronald J. Brachman and Hector J. Levesque. 2022. *Machines like us: toward AI with common sense*. The MIT Press, Cambridge, Massachusetts.

Ruben Branco, António Branco, João António Rodrigues, and João Ricardo Silva. 2021. Shortcutted Commonsense: Data Spuriousness in Deep Learning of Commonsense Reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1521, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*.

Joanna J. Bryson. 2010. Robots should be slaves. In *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, volume 8, pages 63–74. John Benjamins Pub. Company. Publisher: John Benjamins Amsterdam.

Cameron Buckner. 2020. Understanding adversarial examples requires a theory of artefacts for deep learning. *Nature Machine Intelligence*, 2(12):731–736.

Daniel Casasanto and Roberto Bottini. 2014. Mirror reading can reverse the flow of time. *Journal of Experimental Psychology: General*, 143(2):473–479.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W. Black. 2021. Grounding 'Grounding' in NLP. *arXiv:2106.02192 [cs]*. ArXiv: 2106.02192.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv:2204.02311 [cs]*. ArXiv: 2204.02311.

Paul M Churchland and Patricia Smith Churchland. 1990. Could a Machine Think?

Colin Conwell and Tomer Ullman. 2022. Testing Relational Understanding in Text-Guided Image Generation. ArXiv:2208.00005 [cs].

Chaz Firestone. 2020. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal Abstractions of Neural Networks. *arXiv:2106.02997 [cs]*. ArXiv: 2106.02997.

Adele E. Goldberg. 2015. Compositionality. In *The Routledge handbook of semantics*, pages 419–433. Routledge.

John-Stewart Gordon and David J. Gunkel. 2021. Moral Status and Intelligent Robots. *The Southern Journal of Philosophy*, n/a(n/a).

H. P. (Herbert Paul) Grice. 1989. *Studies in the way of words*. Cambridge, Mass. ; London : Harvard University Press.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346. ISBN: 0167-2789 Publisher: Elsevier.

Lennart Heim. 2022. Estimating PaLM's training cost.

Dan Hendrycks, Mantas Mazeika, Andy Zou, Sahil Patel, Christine Zhu, Jesus Navarro, Dawn Song, Bo Li, and Jacob Steinhardt. 2021. What Would Jiminy Cricket Do? Towards Agents That Behave Morally. *arXiv:2110.13136 [cs]*. ArXiv: 2110.13136.

Felix Hill, Andrew Lampinen, Rosalia Schneider, Stephen Clark, Matthew Botvinick, James L. McClelland, and Adam Santoro. 2020. Environmental drivers of systematicity and generalization in a situated agent. *arXiv:1910.00571 [cs]*. ArXiv: 1910.00571.

Douglas R. Hofstadter and Emmanuel Sander. 2013. *Surfaces and essences: analogy as the fuel and fire of thinking*. Basic Books, New York. OCLC: ocn632081685.

Sarah Blaffer Hrdy. 2009. *Mothers and others: the evolutionary origins of mutual understanding*. Belknap Press of Harvard University Press, Cambridge, Mass.

Ray Jackendoff, Neil Cohn, and Bill Griffith. 2012. *A User's Guide to Thought and Meaning*. Oxford University Press, Incorporated, Oxford, UNITED KINGDOM.

Andrew John and Micah Musser. 2022. AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress? Technical report, Center for Security and Emerging Technology.

W. Jeffrey Johnston and Stefano Fusi. 2022. Abstract representations emerge naturally in neural networks trained to perform multiple tasks. Technical report, bioRxiv. Section: New Results Type: article.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv:2001.08361 [cs, stat]*. ArXiv: 2001.08361.

Nikolina Koleva, Martín Villalba, Maria Staudte, and Alexander Koller. 2015. The Impact of Listener Gaze on Predicting Reference Resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 812–817, Beijing, China. Association for Computational Linguistics.

Ray Kurzweil. 2005. *The singularity is near: When humans transcend biology*. Penguin.

Brenden M. Lake and Gregory L. Murphy. 2021. Word meaning in minds and machines. *Psychological Review*.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago Press, Chicago.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit Representations of Meaning in Neural Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online. Association for Computational Linguistics.

Tal Linzen. 2020. How Can We Accelerate Progress Towards Human-like Linguistic Generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Charles Lovering and Ellie Pavlick. 2022. Unit Testing for Concepts in Neural Networks. ArXiv:2208.10244 [cs].

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ.

Christopher D. Manning. 2022. Human Language Understanding & Reasoning. *Daedalus*, 151(2):127–138.

Gary Marcus, Ernest Davis, and Scott Aaronson. 2022. A very preliminary analysis of DALL-E 2. Technical Report arXiv:2204.13807, arXiv. ArXiv:2204.13807 [cs] type: article.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. *arXiv:1902.01007 [cs]*. ArXiv: 1902.01007.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2021. How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN. *arXiv:2111.09509 [cs]*. ArXiv: 2111.09509.

D. McDermott. 1976. Artificial intelligence meets natural stupidity. *SGAR*.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021a. Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand? *arXiv:2104.10809 [cs]*. ArXiv: 2104.10809.

William Merrill, Vivek Ramanujan, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021b. Effects of Parameter Norm Growth During Transformer Training: Inductive Bias from Gradient Descent. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1766–1781, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Julian Michael. 2020. To Dissect an Octopus: Making Sense of the Form/Meaning Debate.

Julian Michael, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang, and Samuel R. Bowman. 2022. What Do NLP Researchers Believe? Results of the NLP Community Metasurvey. ArXiv:2208.12852 [cs].

Ruth Garrett Millikan. 2017. *Beyond concepts: unicepts, language, and natural information*, first edition edition. Oxford University Press, Oxford, United Kingdom. OCLC: ocn989742549.

Marvin Minsky. 1995. Annual Minsky Loebner Prize Revocation Prize 1995 Announcement.

Melanie Mitchell. 2019. *Artificial intelligence: a guide for thinking humans*, first edition. edition. Farrar, Straus and Giroux, New York. Book Title: Artificial intelligence : a guide for thinking humans.

Michael Noukhovitch, Travis LaCroix, Angeliki Lazaridou, and Aaron Courville. 2021. Emergent Communication under Competition. *arXiv:2101.10276 [cs]*. ArXiv: 2101.10276.

Catherine Olsson, Nelson Elhage, Neel Nanda, and Nicholas Joseph. 2022. In-context Learning and Induction Heads. Technical report, Anthropic.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2022. The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. *arXiv:2204.05149 [cs]*. ArXiv: 2204.05149.

Chris Potts. 2022. Could a purely self-supervised Foundation Model achieve grounded language understanding?

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv:2204.06125 [cs]*. ArXiv: 2204.06125.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. *arXiv:2102.12092 [cs]*. ArXiv: 2102.12092.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. 2022. A Generalist Agent. Technical Report arXiv:2205.06175, arXiv. ArXiv:2205.06175 [cs] type: article.

Stuart J. Russell and Peter Norvig. 2021. *Artificial intelligence: a modern approach*, fourth edition edition. Pearson series in artificial intelligence. Pearson, Hoboken.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. Technical Report arXiv:2205.11487, arXiv. ArXiv:2205.11487 [cs] type: article.

Adam Santoro, Andrew Lampinen, Kory Mathewson, Timothy Lillicrap, and David Raposo. 2022. Symbolic Behaviour in Artificial Intelligence. Number: arXiv:2102.03406 arXiv:2102.03406 [cs].

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense Reasoning for Natural Language Processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online. Association for Computational Linguistics.

John R. Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. SCROLLS: Standardized CompaRison Over Long Language Sequences. Technical Report arXiv:2201.03533, arXiv. ArXiv:2201.03533 [cs, stat] type: article.

Murray Shanahan and Melanie Mitchell. 2022. Abstraction for Deep Reinforcement Learning. Number: arXiv:2202.05839 arXiv:2202.05839 [cs].

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image-Chat: Engaging Grounded Conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429, Online. Association for Computational Linguistics.

Linda B. Smith, Swapnaa Jayaraman, Elizabeth Clerkin, and Chen Yu. 2018. The developing infant creates a curriculum for statistical learning. *Trends in cognitive sciences*, 22(4):325–336. Publisher: Elsevier.

Douglas E. Sperry, Linda L. Sperry, and Peggy J. Miller. 2019. Reexamining the Verbal Environments of Children From Different Socioeconomic Backgrounds. *Child Development*, 90(4):1303–1318.

Richard Sutton. 2019. The Bitter Lesson.

Max Tegmark. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*, 1st edition edition. Allen Lane, London.

Neil C. Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso. 2021. Deep Learning's Diminishing Returns: The Cost of Improvement is Becoming Unsustainable. *IEEE Spectrum*, 58(10):50–55. ISBN: 0018-9235 Publisher: IEEE.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent

Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. *arXiv:2201.08239 [cs]*. ArXiv: 2201.08239.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. *arXiv:2204.03162 [cs]*. ArXiv: 2204.03162.

Michael Tomasello. 2003. *Constructing a language*. Harvard university press.

Alan Turing. 1948. Intelligent Machinery.

Alan Turing. 1950. Computing Machinery and Intelligence. *Mind*, 59(236):433.

Joseph Weizenbaum. 1976. *Computer power and human reason : from judgment to calculation*. San Francisco : W. H. Freeman.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020. PuzzLing Machines: A Challenge on Learning From Small Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1241–1254, Online. Association for Computational Linguistics.