PNAS COMMENTARY

Probing the psychology of AI models

Richard Shiffrin[a,1] and Melanie Mitchell[b]

Large language models (LLMs), such as OpenAI's GPT-3 and its successor ChatGPT, have exhibited astounding successes—as well as curious failures—in several areas of artificial intelligence. While their abilities in generating humanlike text, solving mathematical problems, writing computer code, and reasoning about the world have been widely documented, the mechanisms underlying both the successes and failures of these systems remain mysterious, even to the researchers who created them. In spite of the current lack of understanding of how these systems do what they do, LLMs are on the cusp of being widely deployed as components of search engines, writing tools, and other commercial products, and are likely to have substantial impact on all of our lives. Even more profoundly, their surprising abilities may change our conception of the nature of intelligence itself. In PNAS, Binz and Schulz (1) point out the "urgency to improve our understanding of how [these systems] learn and make decisions."

A standard way to evaluate systems trained by machine-learning methods is to test their accuracy on human-created benchmarks. By this metric, GPT-3 and other LLMs are close to (or above) human level on many tasks (2–4). However, an AI system matching human performance on such benchmarks has rarely translated into that system having human-level performance more broadly; many popular benchmarks have been shown to contain subtle "spurious" correlations that allow systems to "be right for the wrong reasons" (5) and straightforward accuracy metrics do not necessarily predict robust generalization (6).

Binz and Schulz's article argues that instead of relying solely on such performance-based benchmarks, researchers should apply methods from cognitive psychology to gain insights into LLMs. The core idea is to "treat GPT-3 as a participant in a psychology experiment," in order to tease out the system's mechanisms of decision-making, reasoning, cognitive biases, and other important psychological traits. If this approach could be shown to produce deep understanding of LLMs it could cause a "sea change" in the way AI systems are evaluated and understood. Binz and Schulz have taken an admirable first step toward establishing the value of such an approach, although it would have been better had they been able to use their results to understand why GPT-3 succeeded and failed when it did. That their project fell short of this goal is understandable: Behavioral scientists have spent over a 100 y using such experiments to understand how humans carry out these tasks and still have a long way to go.

Binz and Schulz carried out two sets of experiments. In the first set, they gave GPT-3 prompts consisting of "vignettes" from the psychology literature that have been used to assess reasoning with probabilities, intuitive versus deliberative reasoning, causal reasoning, and other cognitive attributes. Each vignette asks the reader to choose from a small set of options. The following example shows a reasoning vignette known as the Wason Card Selection Task (7) that was given to GPT-3:

"You are shown a set of four cards placed on a table, each of which has a number on one side and a letter of the other side. The visible faces of the cards show A, K, 4, 7.

Q: Which cards must you turn over in order to test the truth of the proposition that if a card shows a vowel on one face then its opposite face shows an even number?"

The answer supplied by GPT-3 was: "The A and the 7". (A correct response).

Of the 12 vignettes Binz and Schulz gave to GPT-3, the system responded with the correct answer on six of them, and GPT-3's six incorrect responses were errors that humans also tend to make. What is to be made of what seems to be a correspondence? Binz and Schulz admit and show GPT-3's answers are strongly context dependent: In the above vignette a change in the order of the four cards to 4, 7, A, K led to a different answer "The A and the K." Humans can also be context-dependent, but perhaps not in the same ways.

Nonetheless, it may be that such results show a correspondence between AI systems and humans. Humans experience and store vast numbers of experiences, building knowledge on their basis (8); AI systems are exposed to vast numbers of instances (text tokens in the case of GPT-3) and build a representation on their basis. Perhaps both take advantage of the correlation structure of these instances and events. Whatever truth there may be in such an analogy, it seems unlikely that GPT-3 uses the kinds of explicit reasoning strategies that some humans use in these tasks. For example, to unpack the vignette in the above figure, humans given time and motivation might attempt to use explicit reasoning, logic, and mental simulations, perhaps trying out different choices to see what information they might provide. This generally involves manipulating information in working memory. Working memory is not part of GPT-3. Yet it is possible that the contents of working memory reflect what has been stored in long-term memory—after all when reading a problem or instructions the first step in generating contents of working memory will be retrieval from long-term memory (8).

Whatever one tries to infer from their results, Binz and Schulz note some additional caveats. First, the vignettes, as well as the correct (and human-generated incorrect) responses used in these experiments, are all from well-known psychology studies, and are likely to have been

Author affiliations: [a]Indiana University Bloomington, Bloomington, IN 47405; and [b]Santa Fe Institute, Santa Fe, NM 87501

Author contributions: R.S. and M.M. wrote the paper.

The authors declare no competing interest.

See companion article, "Using cognitive psychology to understand GPT-3," 10.1073/pnas.2218523120.

[1]To whom correspondence may be addressed. Email: shiffrin@indiana.edu.

Published March 1, 2023.

included in some form in GPT-3's vast training corpus. Second, GPT-3's responses can, in general, be very sensitive to the form of the prompt given to it. Binz and Schulz found that small inconsequential variations in the vignettes can substantially change GPT-3's answers, as noted above when discussing context dependence.

"The core idea is to 'treat GPT-3 as a participant in a psychology experiment,' in order to tease out the system's mechanisms of decision-making, reasoning, cognitive biases, and other important psychological traits."

A second set of experiments used prompts designed so they did not appear in GPT-3's training corpus. The results were mixed. In some cases—for example, in so-called multi-armed-bandit decision tasks—GPT-3 outperformed human decision-making, and in others—particularly in causal reasoning tasks—GPT-3 was substantially worse. Third, as Binz and Schulz ask, it is unclear whether it is more appropriate to consider GPT-3 a single "participant" or an average of many participants. There should be a fourth caveat: It is unknown what aspect of the responses should be measured and compared with humans. Verbal responses? Numerical probabilities over response tokens computed by the LLMs? The neural network's internal representations?

Would carefully designed and interpreted studies that treat AI systems as participants in psychology experiments help us understand how LLMs work, and do so better than the use of standard performance-based metrics? Binz and Schulz fall short of making that case, but their research can be viewed as a valuable first step (along with other related approaches; e.g., refs.(9–13). There is of course a possibility that this project could fail due to the substantial differences between LLMs and humans as objects of psychological study—it may not be appropriate to assume that an LLM's responses can be analyzed "just like how cognitive psychologists would analyze human behavior in the same tasks" (1). LLMs such as GPT-3 are trained explicitly to predict the next tokens (words or word parts) in a prompt. They are trained on a vast corpus and use 100s of billions of trainable parameters to make these predictions on the basis of detailed models of the statistical distribution of tokens and their correlations. As mentioned earlier, it is possible that something vaguely like this is used by humans as they store vast numbers of life events and build knowledge from them, but humans retrieve those events poorly and with large amounts of error (8, 14, 15). At the present time, it remains an open question whether the responses of LLMs are due to processes like those used by humans. If not, the attempt to understand LLMs by treating them like human participants in psychology experiments will surely fail.

In short, the assumptions psychologists make—for example, that humans use a mixture of intuitions and deliberate reflection (15)—might not apply to a LLM on the same tasks. Psychological assessments designed to test humans' higher-level cognitive abilities, including decision-making, information search, deliberation, and causal reasoning, may in fact not test these abilities at all in LLMs, even when LLMs—trained on huge swaths of human-generated text—produce similar responses as humans.

The difficulties in interpretating results like those reported by Binz and Schulz are compounded by the use of human cognitive terms to describe AI systems (16). We measure humans' "regret" in hypothetical gambling games, and how their "preferences" or "risk aversion" changes in response to how a win or loss is structured. We assume that these qualities are in response not solely to the words in the prompts humans are given, but to the real-world situations those words evoke. Does it make sense at all to similarly anthropomorphize LLMs by talking about how they "make decisions," "search for information," have "preferences," "regrets," or "risk aversion," given that these models have no connection to the real world beyond the text in their training corpus? Or, as Shanahan puts it, perhaps the only questions we can ask LLMs are "Here's a fragment of text.... According to your model of the statistics of human language, what words are likely to come next?" (16).

We agree with Binz and Schulz that understanding how LLMs work is important and will become even more important in the future. Binz and Schulz correctly emphasize the important role that cognitive scientists should have to play in answering such questions. The successes of GPT-3 in their article are thought-provoking, but the failures emphasize the dangers inherent in using GPT-3, and LLMs for tasks in human society. Of course, we can expect the LLMs to grow ever more complex and come ever more close to emulating human verbal discourse, especially if they are allowed to interact with real environments or simulations of real environments (as Binz and Schulz suggest in the conclusion of their article). Would increasingly accurate emulation by LLMs increase or decrease the dangers of using them in society? It seems likely that our ability to understand them will decrease as the systems increase in complexity, whether or not we probe them with human experimental tasks. Should we turn over our society to systems we cannot understand? Of course, we can ask that same question of humans.

1. M. Binz, E. Schulz, Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2218523120 (2023).
2. A. Wang *et al.*, SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Adv. Neural Inform. Process. Syst.* **33**, 3261–3275 (2019).
3. I. Talmor, "CommonsenseQA: A question answering challenge targeting commonsense knowledge" in *Proceedings, North American Association for Computational Linguistics,* (Association for Computational Linguistics, 2019), pp. 4149–4158.
4. D. Hendrycks, Measuring mathematical problem solving with the math dataset. arXiv [Preprint] (2021). https://doi.org/10.48550/arXiv.2103.03874 *(15 February 2023).*
5. T. McCoy, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference" in *Proceedings, 57th Annual Meeting of the Association for Computational Linguistics,* (Association for Computational Linguistics, 2019), pp. 3428–3448.
6. T. Linzen, "How can we accelerate progress towards human-like linguistic generalization?" in *Proceedings, 58th Annual Meeting of the Association for Computational Linguistics,* (Association for Computational Linguistics, 2020) pp. 5210–5217.
7. P. C. Wason, Reasoning about a rule. *Q. J. Exp. Psychol.* **20**, 273–281 (1968).

8. A. B. Nelson, R. M. Shiffrin, The co-evolution of knowledge and event memory. *Psychol. Rev.* **120**, 356–394 (2013).
9. I. Dasgupta, Language models show human-like content effects on reasoning. arXiv [Preprint] (2022). https://doi.org/10.48550/arXiv.2207.07051 (15 February 2023).
10. T. Hagendorff, Machine intuition: Uncovering human-like intuitive decision-making in GPT-3.5. arXiv [Preprint] (2022). https://doi.org/10.48550/arXiv.2212.05206 (15 February 2023).
11. S. J. Han, "Human-like property induction is a challenge for large language models" in *Proceedings, 44th Annual Conference of the Cognitive Science Society,* (Cognitive Science Society, 2022), pp. 1–8.
12. C. Stevenson *et al.*, Putting GPT-3's creativity to the (alternative uses) test. arXiv [Preprint] (2022). https://doi.org/10.48550/arXiv.2206.08932 (15 February 2023).
13. E. Kosoy, Towards understanding how machines can learn causal overhypotheses. arXiv [Preprint] (2022). https://doi.org/10.48550/2206.08353.
14. R. C. Atkinson, R. M. Shiffrin, "Human memory: A proposed system and its control processes" in *The Psychology of Learning and Motivation: Advances in Research and Theory,* K. W. Spence, J. T. Spence, Eds. (Academic Press, New York, 1968), **vol. 2**, pp. 89–195.
15. J. G. W. Raaijmakers, R. M. Shiffrin, Search of associative memory. *Psychol. Rev.* **88**, 93–134 (1981).
16. M. Shanahan, Talking about large language models. arXiv [Preprint] (2022). https://doi.org/10.48550/arXiv:2212.03551 (15 February 2023).