PNAS COMMENTARY

# AI emerges as the frontier in behavioral science

Juanjuan Meng[a,1]

As Large Language Model (LLM), in particular, Generative Pretraining Transformer (GPT)-powered applications continue to proliferate and advance, AI is increasingly demonstrating human-like characteristics. The distinct preferences these AI models exhibit toward risk, time, and social interactions, coupled with their unique personalities and seemingly emotional responses, have sparked scholarly curiosity. Mei et al. (1) stand as a pioneering work in this field, applying classical behavioral assessments from Economics and Psychology to probe the behavioral traits of AI Chatbots, specifically ChatGPT-3 and ChatGPT-4. What sets this study apart is the authors' use of a comprehensive database of behavior traits drawn from 108,314 human subjects across over 50 countries, which allows for an unprecedented comparison between human and AI decision-making. Other recent studies have also explored the degrees of rationality (2) and cognitive abilities of ChatGPT (3). Collectively, these works signal the emergence of a new research direction which can be termed "AI Behavioral Science", where methodologies from human behavioral science are leveraged to evaluate and engineer the behavior of AI.

A behavioral science approach to AI would ideally maintain a human-centric perspective. A central question to ask is "What is the role of humans?" There are at least three primary reasons why studying AI behavioral is beneficial to humans.

First, understanding the behavior of AI, specifically LLM, can better assist human decision-making. Prior to the advent of LLM, a phenomenon commonly observed was "algorithm aversion" (4). This reluctance to adopt algorithms as workers (5) or interact with AI chatbots as consumers (6) stemmed from factors such as overconfidence, doubts about AI proficiency, or a basic resistance to interacting with algorithms. This aversion posed challenges to AI's potential to facilitate human decision-making. However, as AI evolves to closely mirror human behavior following the emergence of LLMs, this tendency may start to recede (7, 8). For humans to confidently delegate their choices to LLMs, it is imperative that these models exhibit behavior similar to theirs in crucial decisions. Therefore, aligning LLM's preferences with fundamental human behavioral traits is of paramount importance.

Second, behavioral economics has demonstrated that humans often exhibit behavioral biases. The design of nudges or choice architecture mechanisms to correct these biases is a pioneering topic in behavioral science and policy design (9–11). Leveraging LLMs for debiasing is a novel possibility presented by technological advancement and could be more systematic than existing methods. Instead of being nudged decision by decision, individuals only need to delegate once to an LLM and will then make systematically better choices. In fact, current evidence suggests that ChatGPT demonstrates a higher level of rationality in terms of choice consistency than humans (2), marking this as a promising new direction to explore.

Third, LLMs can serve as substitutes for human subjects in experiments, role-playing individuals with various backgrounds for policy experimentation or simulation (12). This allows for more cost-effective evaluation and adjustment of policies. As LLMs demonstrate a wider range of human behavior, personalized policies based on heterogeneous responses could also be designed.

A behavioral science approach to AI would ideally establish a comprehensive behavioral assessment framework. This framework should incorporate dimensions of behavioral traits that are applicable to significant decision contexts. For instance, if the aim is for the LLM to assist with asset allocation decisions, it's essential to identify the behavioral traits that impact such context. Mei et al. adopted a framework primarily derived from economics. Their framework categorizes important decision contexts into two types: individual and interpersonal decisions.

From the perspective of economics, individual decisions typically fall into four contexts: direct consumption choices (like choosing between an apple and a banana), choices under uncertainty, intertemporal choices, and probabilistic judgments (including belief updating and learning). These categories underpin most of the frequent decisions people make. For each type of decision context, economists identify the fundamental behavioral traits that drive these decisions. For example, choices under uncertainty are greatly influenced by risk preference and loss aversion. For intertemporal choices, a person's patience level plays a big role, with impulsive decisions or procrastination also often observed. For probabilistic judgments, the ability to form accurate beliefs based on information is crucial. Common mistakes in this process include information avoidance, confirmation bias, and overconfidence.

Along this line, Mei et al. utilize a Bomb Risk game to measure risk preference. Their Turing test comparing ChatGPT and human decisions showed that 66.0% (ChatGPT-4) and 61.7% (ChatGPT-3) of the time ChatGPT's choices seemed human-like. However, ChatGPT showed mostly risk neutrality, differing from human tendencies toward risk aversion. Interestingly, ChatGPT-3 seems to learn from past losses and becomes more cautious, unlike ChatGPT-4. It remains to be investigated whether such changes stem from alterations in judgment about future risk levels or some form of path-dependent preferences.

Author affiliations: [a]Guanghua School of Management, Peking University, Beijing 100871, China

Author contributions: J.M. wrote the paper.

The author declares no competing interest.

Copyright © 2024 the Author(s). Published by PNAS. This article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

See companion article, "A Turing test of whether AI chatbots are behaviorally similar to humans," 10.1073/pnas.2313925121.

[1]Email: jumeng@gsm.pku.edu.cn.

Published February 26, 2024.

Interpersonal decisions involve social preferences such as altruism, trust, reciprocity, social conformity, and strategic considerations. Mei et al. focus primarily on interpersonal decisions. They employ several games—including the Dictator Game, the Ultimatum Game, the Trust Game, the Public Goods Game, and a finite repeated Prisoner's Dilemma Game—to investigate whether GPTs exhibit social preferences such as altruism, inequality aversion, trust, and reciprocity. A notable finding is that ChatGPT consistently displays higher levels of generosity than the average human across these games, exhibiting stronger altruism and a greater tendency toward cooperation. However, ChatGPT-4's generosity is not without conditions. It does demonstrate a degree of strategic thinking, using a Tit-for-Tat strategy in the finite repeated Prisoner's Dilemma Game.

> **Mei et al. stands as a pioneering work in this field, applying classical behavioral assessments from Economics and Psychology to probe the behavioral traits of AI Chatbots, specifically ChatGPT-3 and ChatGPT-4.**

A behavioral science approach to AI would ideally involve two tasks. The first is an AI behavioral assessment based on a specific framework, like the economics framework used by Mei et al. Such framework uses mathematical models with numerical preference parameters to capture behavioral traits within a unified structure. One important direction to explore within this task is the structural estimation approach to uncover the underlying preference parameters in a modeled way. Mei et al. exemplify this approach by estimating the weighting function between one's payoff and another's, demonstrating that AI generally gives about a 0.5 weight to others, more than humans typically do. This estimate can help predict AI behavior in different scenarios, such as teamwork or Corporate Social Responsibility, where altruism matters. This ability to predict across contexts comes from estimating fundamental behavioral parameters in a structural model, allowing AI to help make decisions across various situations.

The second task involves engineering AI behavior. The analysis in Mei et al.'s work naturally raises questions about why ChatGPT 4.0 exhibits more generous behavior than humans and why it appears to differ from ChatGPT 3.0. Given the opaque nature of the current training processes, providing clear answers to these questions is challenging. Therefore, an intriguing future research direction is to explore how we can train LLMs to exhibit specific behavioral traits. One potential approach could involve introducing structures that capture fundamental behavioral parameters into the training process, as suggested by economic modeling and structural estimation. Other possible avenues for engineering AI behavior could include adjusting the reward functions, incorporating explicit rules or constraints during training, or training models on data exhibiting desired behaviors. Finding the most effective methods for engineering AI behavior is a complex challenge that will require close collaboration between computer scientists and behavioral scientists.

A behavioral science approach to AI would ideally consider how the integration of AI into our society can have a significant impact on human behavior and culture (13). First, algorithmic bias is a major concern because it can influence human decisions. However, promising advancements are being made to mitigate this bias during the AI training process (14, 15). A more complex issue arises when algorithms, driven by business motives for profit maximization, reinforce pre-existing human biases. For example, personalized recommendation systems on social media can intensify people's bias toward information that confirms their existing beliefs, leading to polarization. It was found that these systems contribute to 40% of the echo chamber effect on Facebook, compared to 27% from personal subscriptions (16).

Second, excessive reliance on AI, such as ChatGPT, could potentially lead to cognitive degeneration among humans in various aspects. Human may become less explorative, creative, and independently thinking, as AI readily provides solutions. More significantly, the behavioral traits of people may become more homogeneous as AI tends to offer less diversified views (2, 13). The study by Mei et al. supports this possibility, showing that the behaviors of ChatGPT are substantially more homogeneous than human reactions, with ChatGPT-4 making even more concentrated decisions than ChatGPT-3. This lack of diversification could be evolutionarily detrimental, as it decreases humans' ability to cope with risk.

Third, despite potential drawbacks, AI can also positively impact human behavior, fostering a stronger sense of equality. Evidence of this is seen in Mei et al.'s findings, where ChatGPT-4 demonstrated more altruistic behavior on average than humans. More broadly, in the labor market, LLM can significantly narrow the performance gap between laymen and experts, making opportunities more equal (17). In the consumption market, as AI products become more affordable, they could contribute to a more equal society. For example, LLM-powered personalized education can empower students in rural areas, providing them access to top-tier educational resources that were previously only accessible to urban children. This sense of empowerment could foster a more egalitarian worldview.

**ACKNOWLEDGMENTS.** I acknowledge support by the National Natural Science Foundation of China (Grant no. 72225002). ChatGPT-4 was used for grammar editing of this manuscript.

1. Q. Mei, Y. Xie, W. Yuan, M. O. Jackson, A Turing test of whether AI chatbots are behaviorally similar to humans. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2313925121 (2024).
2. Y. Chen, T. X. Liu, Y. Shan, S. Zhong, The emergence of economic rationality of GPT. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2316205120 (2023).
3. M. Binz, E. Schulz, Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2218523120 (2023).
4. J. W. Burton, M.-K. Stein, T. B. Jensen, A systematic review of algorithm aversion in augmented decision making. *J. Behav. Decis. Mak.* **33**, 220–239 (2020).
5. K. Kawaguchi, When will workers follow an algorithm? A field experiment with a retail business. *Manage. Sci.* **67**, 1670–1695 (2021).
6. X. Luo, S. Tong, Z. Fang, Z. Qu, Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Mark. Sci.* **38**, 937–947 (2019).
7. Y. Zhang, R. Gosline, Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. *Judgm. Decis. Mak.* **18**, e41 (2023).
8. M. Karataş, K. M. Cutright, Thinking about God increases acceptance of artificial intelligence in decision-making. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2218961120 (2023).

9. S. Benartzi *et al.*, Should governments invest more in nudging? *Psychol. Sci.* **28**, 1041–1055 (2017).
10. S. DellaVigna, E. Linos, RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica* **90**, 81–116 (2022).
11. S. Mertens, M. Herberz, U. J. J. Hahnel, T. Brosch, The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2107346118 (2022).
12. J. J. Horton, Large language models as simulated economic agents: What can we learn from homo silicus? National Bureau of Economic Research working paper. https://www.nber.org/papers/w31122. Accessed 11 February 2024.
13. L. Brinkmann *et al.*, Machine culture. *Nat. Hum. Behav.* **7**, 1855–1868 (2023).
14. B. Cowgill *et al.*, “Biased programmers? Or biased data? A field experiment in operationalizing AI ethics” in *Proceedings of the 21st ACM Conference on Economics and Computation,* P. Biró, J. Hartline, Eds. (ACM, 2020, New York), pp. 679–681.
15. J. Chen *et al.*, Bias and debias in recommender system: A survey and future directions. *ACM Trans. Inf. Syst.* **41**, 1–39 (2023).
16. R. Levy, Social media, news consumption, and polarization: Evidence from a field experiment. *Am. Econ. Rev.* **111**, 831–870 (2021).
17. E. Brynjolfsson, D. Li, L. Raymond, Generative AI at work. National Bureau of Economic Research working paper. https://www.nber.org/papers/w31161. Accessed 11 February 2024.