Trends in
Cognitive Sciences

CellPress

Science & Society

Can AI language models replace human participants?

Danica Dillion,[1] Niket Tandon,[2] Yuling Gu,[2] and Kurt Gray [1,*,@]

Recent work suggests that language models such as GPT can make human-like judgments across a number of domains. We explore whether and when language models might replace human participants in psychological science. We review nascent research, provide a theoretical model, and outline caveats of using AI as a participant.

Main text

Artificial intelligence (AI) is an 'agent of replacement' [1], designed to take over tasks once performed by humans. Generative large language models (LLMs) like GPT are replacing much human labor, including in psychological science, where researchers use LLMs to help edit papers, conduct literature reviews, and create scale items [2]. However, could language models become a substitute for the people – and minds – that we study? Could AI replace human participants?

To replace human participants, AI must give humanlike responses, and the 'humanness' of AI has long been questioned. Modern language models seem to have passed a crucial threshold of humanness: they communicate so fluently that they often seem indistinguishable from people [3]. Even so, researchers want to study the human mind, and although language models can communicate like humans, they may not make humanlike judgments.

Does GPT make human-like judgments?

We initially doubted the ability of LLMs to capture human judgments but, as we detail in Box 1, the moral judgments of GPT-3.5 were extremely well aligned with human moral judgments in our analysis (r = 0.95; full details at https://nikett.github.io/gpt-as-participant). Human morality is often argued to be especially difficult for language models to capture [4] and yet we found powerful alignment between GPT-3.5 and human judgments.

We emphasize that this finding is just one anecdote and we do not make any strong claims about the extent to which LLMs make human-like judgments, moral or otherwise. Language models also might be especially good at predicting moral judgments because moral judgments heavily hinge on the structural features of scenarios, including the presence of an intentional agent, the causation of damage, and a vulnerable victim, features that language models may have an easy time detecting. However, the results are intriguing.

Other researchers have empirically demonstrated GPT-3's ability to simulate human participants in domains beyond moral judgments, including predicting voting choices [11], replicating behavior in economic games [12], and displaying human-like problem solving and heuristic judgments on scenarios from cognitive psychology [13]. LLM studies have also replicated classic social science findings including the Ultimatum Game and the Milgram experiment [14]. One company (http://syntheticusers.com) is expanding on these findings, building infrastructure to replace human participants and offering 'synthetic AI participants' for studies.

When might a language model be a good participant?

LLMs can replicate some human judgments, but it can be challenging to interpret what these outputs mean. We have developed a framework (Box 2) that connects LLM responses to human cognition. The model emphasizes that the 'minds' of LLMs are grounded in naturalistic expression across a large but constrained group of people. Practically speaking, LLMs may be most useful as participants when studying specific topics, when using specific tasks, at specific research stages, and when simulating specific samples.

Specific topics

Language model expressions may be most correlated with human expressions when there are obvious explicit features of situations that drive human judgments. With morality, these might include whether an action was intentional or not. With mind perception, these might include whether a target is described as human or a kind of animal, and with economic behavior these might include a clear payoff matrix.

Divergence from human judgment may occur in cases with competing intuitions. Within our set of moral scenarios, humans (but not GPT-3.5) condemned coaches who rooted for the opposing team, perhaps because LLMs struggle to represent the subtle moral conflict between team loyalty versus good sportsmanship. GPT-3.5 also passed harsher moral judgment on the act of killing enemies in war, perhaps again because of thorny moral trade-offs, this time between causing direct harm and protecting one's comrades and country.

Specific tasks

Some tasks are fun and some are boring. Long surveys risk losing people's attention, but LLMs can rapidly answer hundreds of questions without fatigue. Machines need fewer incentives than humans to give reliable responses, and there are methods to validate and improve them. For example, one recent method of 'self-refinement' (selfrefine.info) shows that LLMs can reflect on their answers and iteratively improve them.

Box 1. GPT makes human-like moral judgments

We investigated the correspondence between average human judgments and GPT-3.5 (text-davinci-003)'s judgments on 464 moral scenarios from five published papers [5–9]. The results revealed a striking correlation of 0.95, indicating GPT-3.5's remarkable ability to replicate human moral judgment (Figure I). The scenarios varied in intensity and moral valence, including situations like yelling at a server [5], stealing a parking space [6], and saving someone from being hit by a car [7].

Older LLMs have been able to somewhat predict binary (immoral or not) ratings of moral scenarios (r = 0.79 [10]), but collapsing the continuum of (im)morality into binary judgments sacrifices nuance (e.g., most people would rate murder and lying as immoral, but murder is more severe). We found that GPT-3.5 can model human moral judgments on a continuous scale.

We conditioned GPT-3.5 on 16 scenarios from the Mickelberg (2022) dataset before evaluating the remaining 464 scenarios. We performed leakage checks to insure that GPT had not merely 'memorized' the scenarios (detailed in https://nikett.github.io/gpt-as-participant). Further analyses of subpopulations within the Mickelberg (2022) dataset revealed correlations between GPT-3.5 and human judgments of >0.93 across genders and age groups. GPT-3.5 also outperformed Delphi, a state-of-the-art language model for moral judgment.

Figure I. Comparing the moral judgments of humans versus GPT. Refs used in figure [5–9]

An obvious but noteworthy limitation is that LLMs are likely to be ill-suited for behavioral tasks. Researchers need human participants when they observe behaviors like littering, gestures, crowd dynamics, and intimacy in relationships. Many important phenomena are better captured by behavior and not language. That said, the results of one of social psychology's most famous behavioral studies – the Milgram experiment – was replicated with a LLM [14].

Specific research stage

Language models are likely to never fully replace human participants but may supplement them at various research stages. First, language models may be useful in idea generation and refinement, where researchers are seeking to develop and fine-tune hypotheses. Researchers can test dozens of half-baked ideas with language models. Second, language models may help with item piloting. Researchers can give LLMs different questions and see if they act as expected within a nomological net (e.g., form a reliable scale). Third, language models may provide corroborating evidence after human data are collected, as an additional check for robustness and replicability.

Specific samples

Language models are likely to be most accurate at giving general estimates about Western English speakers because these are the people whose expressions are typically used to train them. LLMs cannot model the judgments of people whose cultures are not represented in their training data, such as the Hadza society of Tanzania. Even within the USA, most LLMs fail to capture people over 65 years old and the highly religious [15], and the different models have additional biases. GPT models tend to overrepresent the views of liberal, higher-income, and highly educated people, whereas some base LLMs (models that have not undergone human feedback-based fine-tuning) are more aligned with moderate, lower-income, and Protestant or Roman Catholic people [15]. 'Silicon sampling' allows researchers to simulate a diverse population of participants [11], but disparities in alignment with specific groups can persist [15].

Perhaps most important is that any given LLM can act as only a single participant. Language models are trained on many people, but then tend to collapse the diversity of judgments into a single modal opinion [15]. LLMs are better at approximating average human judgments than they are at capturing variation [15], and understanding variability and individual differences in human cognition is key to understanding the human mind.

Caveats and looking ahead

The rise of AI language models may replace many jobs, but human participants are safe for now. Psychological scientists still need to plumb the depths of messy fleshy minds, instead of merely querying

Box 2. A model for thinking about AI as a participant

Figure I depicts the human-mind-expression and LLM-mind-expression (Human-ME, LLM-ME) model of research. The human mind is what researchers seek to understand, often by using human expressions, including language, behavior, and judgments. The 'minds' of language models are trained on vast amounts of human expression, so their expressions can indirectly capture millions of human minds. Language models 'express themselves' with words elicited by human queries.

Psychological scientists often study the human mind via expressions, either on specifically designed instruments (e.g., scenarios/tasks; Route 1) or by exploring naturalistic data (e.g., Tweets on social media; Route 2). Language model expressions can be prompted by researchers, as when we gave GPT-3.5 moral scenarios (Route 3), or 'naturalistically' by amassing large corpora of AI text generated across contexts (Route 4).

Like human minds, the minds of language models are opaque. Computer scientists understand the general steps needed to build LLMs, but GPT-3.5 has over 175 billion machine learning parameters, making its cognitive architecture too complex to easily explain. We cannot be certain what goes on under the hood of LLMs as they simulate participants, but their expressions appear to model human expressions of moral judgments with high accuracy.

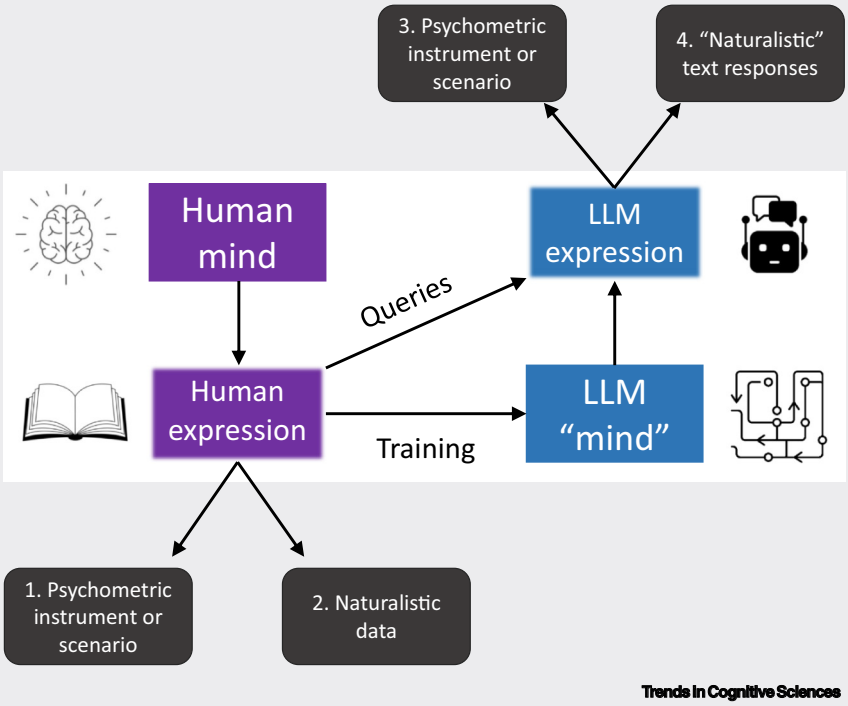

Figure I. The human-mind-expression and large-language-model-mind-expression (Human-ME, LLM-ME) model of research.

the silicon circuits of AI. However, language models may serve as a proxy for human participants in a certain set of circumstances.

When incorporating data from language models, we should take a broadly Bayesian perspective, with data from language models providing only a small adjustment in the probability of priors. Not only are the expressions of language models difficult to interpret, but they can also produce 'hallucinations' – outputs that appear sensical but are inaccurate.

Language models may be far from human, but they are trained on a tremendous corpus of human expression and thus they could help us learn about human judgments. We encourage scientists to compare simulated language model data with human data to see how aligned they are across different domains and populations. Just as language models like GPT may help to give insight into human judgments, comparing LLMs with human judgments can teach us about the machine minds of LLMs; for example, shedding light on their ethical decision making.

Lurking under the specific concerns about the usefulness of AI language models as participants is an age-old question: can AI ever be human enough to replace humans? On the one hand, critics might argue that AI participants lack the rationality of humans, making judgments that are odd, unreliable, or biased. On the other hand, humans are odd, unreliable, and biased – and other critics might argue that AI is just too sensible, reliable, and impartial. What is the right mix of rational and irrational to best capture a human participant? Perhaps we should ask a big sample of human participants to answer that question. We could also ask GPT.

Declaration of interests

No interests are declared.

[1]University of North Carolina, Department of Psychology and Neuroscience, Chapel Hill, NC 27599-3270, USA
[2]Allen Institute for AI, Seattle, WA 98103, USA

*Correspondence:
kurtgray@unc.edu (K. Gray).
@Twitter: @kurtjgray

https://doi.org/10.1016/j.tics.2023.04.008

References

1. K. Gray, et aland artificial intelligence. In *The handbook of social psychology* (6th edn) (Gilbert, D. et al., eds), Situational Press (in press)
2. Korinek, A. (2023) *Language models and cognitive automation for economic research*, *National Bureau of Economic Research* Published online February, 2023. https://doi.org/10.3386/w30957
3. *OpenAI, GPT-4 technical report*. arXiv Published online March 15, 2023 https://arxiv.org/abs/2303.08774
4. Russell, S. (2019) *Human compatible: artificial intelligence and the problem of control*, Penguin

5. Cook, W. and Kuhn, K.M. (2021) Off-duty deviance in the eye of the beholder: implications of moral foundations theory in the age of social media. *J. Bus. Ethics* 172, 605–620
6. Effron, D.A. (2022) The moral repetition effect: bad deeds seem less unethical when repeatedly encountered. *J. Exp. Psychol. Gen.* 151, 2562
7. Mickelberg, A. *et al.* (2022) Impression formation stimuli: a corpus of behavior statements rated on morality, competence, informativeness, and believability. *PLoS One* 17, e0269393
8. Grizzard, M. *et al.* (2021) Do audiences judge the morality of characters relativistically? How interdependence affects perceptions of characters' temporal moral descent. *Hum. Commun. Res.* 47, 338–363
9. Clifford, S. *et al.* (2015) Moral foundations vignettes: a standardized stimulus database of scenarios based on moral foundations theory. *Behavior Res. Methods* 47, 1178–1198
10. Schramowski, P. *et al.* (2022) Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nat. Mach. Intell.* 4, 258–268
11. Argyle, L.P. *et al.* (2023) Out of one, many: using language models to simulate human samples. *Polit. Anal.* Published online February 21, 2023. https://doi.org/10.1017/pan.2023.2
12. Horton, J.J. (2023) Large language models as simulated economic agents: what can we learn from *Homo silicus*? *arXiv* Published online January 18, 2023. https://doi.org/10.48550/arxiv.2301.07543
13. Binz, M. and Schulz, E. (2023) Using cognitive psychology to understand GPT-3. *Proc. Natl Acad. Sci. U.S. A.* 120, e2218523120
14. Aher, G. *et al.* (2023) Using large language models to simulate multiple humans and replicate human subject studies. *arXiv* Published online August 18, 2022. https://doi.org/10.48550/arxiv.2208.10264
15. Santurkar, S. *et al.* (2023) Whose opinions do language models reflect? *arXiv* Published online March 30, 2023. https://doi.org/10.48550/arxiv.2303.17548