AI Language Models Cannot Replace Human Research Participants

Jacqueline Harding (Stanford University, Center for AI Safety), William D'Alessandro (Munich Center for Mathematical Philosophy, Center for AI Safety), N. G. Laskowski (University of Maryland at College Park, Center for AI Safety), Robert Long (Center for AI Safety)[1]

Forthcoming in *AI & Society*

Please cite published draft when available

Generative artificial intelligence has the potential to transform many of our projects, from art and education to business and security. AI's promise in these areas comes largely from its ability to perform complex cognitive tasks more cheaply, quickly and reliably than humans.

Given such capabilities, AI is also likely to have splashy effects on scientific research. Machine learning techniques have already been used to speed up the search for useful drugs, help prove mathematical theorems, refine climate models, and parse enormous datasets from space telescopes and particle accelerators. While large language models (LLMs) and other generative AI tools haven't yet taken center stage, their research potential will undoubtedly grow as context windows expand, reasoning abilities improve, and multimodal proficiencies multiply.

Other potential scientific uses of generative AI, however, are relatively speculative, and require more critical scrutiny. Our plan here is to wax curmudgeonly about one such proposal, pertaining to LLMs and moral psychology research.

The proposal starts from the observation that, having been trained on trillions of internet text tokens, contemporary language models are skilled imitators of human linguistic behavior. LLM responses closely align with an average person's on a variety of prompts; current models have even replicated classical phenomena from psychology, like Hsee's less-is-better effect, and behavioral economics, like Harsanyi's ultimatum games. In light of their speed, ease of recruitment and talents at mimicry, it's natural to wonder what contributions LLMs might make to the human sciences.

Dillion et al.'s (2023) recent "Can AI language models replace human participants?" raises a sharp version of this question. The authors focus on moral psychology, noting that GPT-3.5 (text-davinci-003) produces judgments about a variety of moral scenarios which correlate strongly with average human ratings. Such correlations are impressive, and they deserve further study across a wider range of models and prompts. But, as we'll argue, they don't underwrite any interesting degree of replacement of humans by language models.

Dillion et al. themselves propose three concrete applications of LLMs in moral psychology research: (1) helping generate and refine research hypotheses, (2) piloting test items, and (3)

[1] Authors listed alphabetically. Authors contributed equally.

corroborating data gathered from human subjects. These proposals have some plausibility. However, to the extent that they're plausible, they offer little support for the prospect of human replacement.

Let us elaborate. The authors' proposal 1, generating and refining hypotheses, concerns a stage of research which doesn't inherently involve human participants, so there's no question of replacement to begin with. As for proposal 2, LLMs are unsuitable for important aspects of item piloting. Consider the need to determine whether participants may misinterpret or struggle with a given test item. Because language models are exemplary text-processors by design, they won't accurately model human participants' difficulties with comprehension, reasoning and the like. So humans will still be needed to assess these factors. Proposal 3 calls for language model outputs to serve as comparison points to ordinary experimental data. While an interesting suggestion, this presupposes that LLMs haven't replaced humans as primary research subjects. Dillion et al.'s concrete suggestions therefore seem to involve relatively modest kinds of supplementation, not replacement.

Apart from these three proposals, Dillion et al. also hint at a larger role for LLMs, as reflected in their paper's title. The authors ask whether AI models might "…become a substitute for the people—and minds—that [moral psychologists] study" (p.1). They then claim that "To replace human participants, AI must give humanlike responses" (ibid). Finally, they note that "Recent work suggests…[LLMs] can make human-like judgments" (ibid). This all amounts to an apparently optimistic, or at least open-minded, appraisal of a strong replacement thesis.

What would the motivation for such a thesis look like? Dillion et al. suggest picturesquely that a successful model of a large corpus of human text will "indirectly capture millions of human minds" (p.3). The output of such a model can then be thought of as an expression of a "modal opinion" (p.2) of the captured minds. Of course—as Dillion et al. are quick to observe—the data on which language models are trained was produced by a specific subpopulation of humans, meaning claims made on the basis of the model's representativeness must be carefully circumscribed. Current methods for fine-tuning LLM performance, such as reinforcement learning with human feedback, further exacerbate this issue.

Assume, however, that an LLM *can* output relatively accurate modal opinions for some population. Suppose we present this model with a novel moral vignette for which we have no human data, and its output is intuitively surprising. Is this strong evidence that some population of humans would form that judgment? Or should we suspect that the model has given a non-humanlike response, perhaps because it's latched onto some unconsidered aspect of the prompt, or because the vignette is out of distribution for the model? In scenarios like this, the informativeness of the LLM's output is impossible to assess without doing further confirmatory work with human participants.

One might try to steer a middle course between the replacement fan's optimism and our pessimism by suggesting that evidence from LLMs, while useful enough to take the place of some human data, is nevertheless defeasible and relatively weak on its own. Dillion et al. may be espousing a version of this idea when they recommend taking "a broadly Bayesian perspective, with data from language models providing only a small adjustment in the probability of priors" (p.3).

We have no quarrel with the idea of using LLMs for small Bayesian updates. If GPT-4 classifies controlled forest burns as morally good, say, this gives a bit of reassurance that (some) humans would judge similarly. The problem with this suggestion is that many information sources can provide small Bayesian updates without thereby qualifying as blue-ribbon experimental data; not all useful evidence is admissible scientific evidence. So, Dillion et al.'s modesty is appropriate, but it doesn't significantly strengthen the case for replacement.

Here's a final possible defense of replacement optimism. At present, LLMs are relatively immature, and we haven't yet adequately probed the nature and extent of their ability to simulate human moral judgments. But these factors will improve over time. Conceivably, we'll be so confident about a future GPT-n's accuracy (for some populations, in some domains of interest) that we can cut humans out of the experimental process.

But this response appears to rest on a mistaken assumption about the stability of our judgments over time. Given their training data, language models offer a snapshot of average moral opinion over some fixed past period. But changing world events, personal experiences and social developments mean that moral views are always in flux—and major shifts can happen rapidly, as with American attitudes toward LGBT issues in the 2000s. Consequently, frequent and careful work with human participants will always be an integral part of moral psychology research.[2]

While LLMs will undoubtedly prove useful to scientists, it's unlikely that they can supplant human research participants in any significant way. We remain optimistic about a broader range of AI research applications in psychology and elsewhere. In general, the possibility of using LLMs to model or simulate human behavior presents rich possibilities and questions. The relevant conversations at the intersection of AI, philosophy of science, moral psychology and ethics are only beginning.

On behalf of all authors, the corresponding author states that there is no conflict of interest.

[2] The idea that human moral development is constantly in flux has been espoused by several prominent moral philosophers, including Rawls (1974: 289), Kagan (1998: 16), and Scanlon (1998: 361), the latter of whom writes, "Working out the terms of moral justification is an unending task". See also Laskowski (2019) for precisification and extended defense of this idea.

References

Dillion, D. et al. (2023). “Can AI language models replace human participants?”. *Trends in Cognitive Sciences* 2438, 1-4. DOI: 10.1016/j.tics.2023.04.008.

Kagan, S. (1998). *Normative Ethics*. Boulder: Westview Press.

Laskowski, N. (2018a). “Epistemic Modesty in Ethics”. *Philosophical Studies* 175(7): 1577–96.

Rawls, J. (1974). “The Independence of Moral Theory”. Proceedings and Addresses of the American Philosophical Association, 48, 5–22.

Scanlon, T. (1998). *What We Owe to Each Other*. Cambridge: Harvard University Press.