SOCIAL SCIENCE

AI and the transformation of social science research

Careful bias management and data fidelity are key

By **Igor Grossmann[1,2], Matthew Feinberg[3], Dawn C. Parker[2,4], Nicholas A. Christakis[5], Philip E. Tetlock[6], William A. Cunningham[7,8,9]**

Advances in artificial intelligence (AI), particularly large language models (LLMs), are substantially affecting social science research. These transformer-based machine-learning models pretrained on vast amounts of text data are increasingly capable of simulating human-like responses and behaviors (*1*, *2*), offering opportunities to test theories and hypotheses about human behavior at great scale and speed. This presents urgent challenges: How can social science research practices be adapted, even reinvented, to harness the power of foundational AI? And how can this be done while ensuring transparent and replicable research?

Social sciences rely on a range of methods, including questionnaires, behavioral tests, mixed-method analyses of semi-structured responses, agent-based modeling (ABM), observational studies, and experiments. The common goal is to obtain a generalized representation of characteristics of individuals, groups, cultures, and their dynamics (*2*). With the advent of advanced AI systems, the landscape of data collection in social sciences may shift. LLMs take advantage of deep learning to capture complex relationships within language. Such language literacy capabilities in processing, generating, and interacting with human language in a contextually aware and semantically accurate fashion (*1*) represent a major shift from previous AI approaches, which often struggled with such nuanced aspects of language as irony, metaphor, or emotional tone. With proper conditioning (*3*), LLMs can more accurately simulate human behavioral responses in social science research.

LLMs may supplant human participants for data collection. For example, LLMs have already demonstrated their ability to generate realistic survey responses concerning consumer behavior (*2*). Although opinions on the feasibility of this application vary, at a minimum, studies that use simulated participants could be used to generate hypotheses that could then be confirmed in human populations (*3*, *4*). The success of this approach depends on algorithmic fidelity of the trained data (*3*), transparency in model training, prompt engineering, and benchmark selection.

Why is this scenario plausible? Pretrained on massive datasets, advanced AI models can represent a vast array of human experiences and perspectives, possibly giving them a higher degree of freedom to generate diverse responses than that of conventional human participant methods, which can help to reduce generalizability concerns in research (*2*). LLMs can also generate responses across a wider range of parameters than human participants because of pragmatic concerns of limited attention span, response bias, or habituation among humans, providing a less biased view of underlying latent dimensions. This makes them especially useful in high-risk projects for which traditional data collection is impractical, allowing for the testing of interventions in simulated populations before real-world implementation.

LLMs could be used as surrogates in other ways. They have the potential to enhance policy analysis by reproducing the views of different theoretical or ideological schools of thought. For example, LLMs could be trained to capture nuances of complex debates, such as concerning the stability and reliability of nuclear deterrence in the face of human and technical factors (*5*). LLMs could be trained to capture varied perspectives, including evaluating "what-if" scenarios that nearly occurred, such as the Cuban Missile Crisis in 1962, and providing assessments of how plausible these scenarios were. Once LLMs can pass the Ideological Turing Test—meaning that they can accurately represent opposing viewpoints in a way indistinguishable from real humans—researchers can use them to generate future scenarios. Future LLMs, appropriately trained (*3*), may thus out-perform humans on analytic tasks such as synthesizing clashing views to generate superior forecasts and policy prescriptions.

AI could also fill the role of a "confederate" (controlled experimental partner) in social interaction research involving individuals or groups (*6*), potentially as components to agent-based simulations. An LLM-ABM hybrid could use LLM to derive empirically based rules of social decision-making or behavior to simulate social interactions of individuals with specific characteristics and beliefs (*4*). This approach could explore how agents with these particular characteristics influence subsequent interaction with humans, informing broader social science questions such as how misinformation spreads throughout social networks (*7*).

Such investigations raise questions about the limits of LLMs as human cognition and decision models. Can we "nudge" an LLM by asking it to assess the quality of a news item before sharing, replicating research with humans (*7*)? If so, could we use the integrated LLM-ABM model to identify interventions that would reduce the spread of misinformation through social networks? Generally, if LLM-ABMs can provide new insights on how human agents choose to share information, cooperate and compete in social dilemmas, and conform with social norms, they can provide valuable insights into both the underlying mechanisms governing human behavior and social dynamics (*8*) with higher fidelity than has been possible with previous human decision models.

Incorporating LLMs into ABMs introduces new challenges because of their differing operational principles. Whereas LLMs generate and interpret language according to statistical patterns derived from vast linguistic data, traditional ABMs operate on the basis of predefined formal rules (*9*) that can be generated by using real-world linguistic and other qualitative data. New ABM design will be needed to take advantage of LLMs' capability to simulate performance on questionnaires, behavior in ill-defined situations, or open-ended responses (*2*). By creating realistic initial populations for ABMs, LLMs can model subjects' latent cognitive or affective states, surpassing traditional researchers' capacity and opening doors for future theory generation.

LLMs' potential future benefits include creating samples as diverse as the cultural products (*2*, *3*) on which the models were trained, offering a more accurate

[1]Department of Psychology, University of Waterloo, Waterloo, ON, Canada. [2]Waterloo Institute for Complexity and Innovation, University of Waterloo, Waterloo, ON, Canada. [3]Rotman School of Management, University of Toronto, Toronto, ON, Canada. [4]School of Planning, University of Waterloo, Waterloo, ON, Canada. [5]Yale Institute for Network Science, Yale University, New Haven, CT, USA. [6]Wharton School of Business, University of Pennsylvania, Philadelphia, PA, USA. [7]Department of Psychology, University of Toronto, Toronto, ON, Canada. [8]Vector Institute, Toronto, ON, Canada. [9]Schwartz Reisman Institute for Technology and Society, University of Toronto, Toronto, ON, Canada. Email: igrossma@uwaterloo.ca

portrayal of human behavior and social dynamics than those from conventional methods that rely on typically less heterogenous and representative convenience samples (*2*). Because of their population-scale calibration data, LLMs could help address common challenges in social science research that can lead to biased models, including generalizability and self-selection concerns (*2*).

Effective AI-assisted research will depend on the AI being able to accurately mirror the perspectives of diverse demographic groups. Pretrained models from linguistic cultural products are known to capture sociocultural biases present in society (*2*, *10*). When biases are recognized, a key question is their provenance: Do they correctly reflect the populations, or are they artifacts of model construction (*11*)? Model construction bias may result from incorrect or invalid choices throughout the design and development pipeline (for example, choosing constructs that are differentially valid across demographic groups, curating datasets that lack diversity or that encode biases of certain human annotators, or selecting models that fail to capture specific patterns pertinent to minorities) or because of existing societal disparities (*2*).

The scientist-humanist dilemma emerges as a key issue: Although scientists aim to study "pure" LLMs with embedded sociocultural biases to simulate human behavior and trace its cultural evolution (*2*), ethical constraints require engineers to protect LLMs from these very biases. Already, LLM engineers have been fine-tuning pretrained models for the world that "should be" (*12*) rather than the world that is, and such efforts to mitigate biases in AI training (*2*, *13*) may thus undermine the validity of AI-assisted social science research. The proprietary "black box" nature of LLM training challenges the ability of researchers to evaluate underlying mechanisms and replicate findings. To address this, advocating for open-source LLMs, access to pretrained but not fine-tuned models for scientific research, and transparent methodologies (such as BLOOM, Cerebras-GPT, or LLaMA) are essential for ensuring reliable and credible AI-driven research (*2*).

Overall, researchers will need to establish guidelines for the ethical use of LLMs in research, addressing concerns related to data privacy, algorithmic fairness [versus monoculture (*2*)], environmental costs (*2*, *13*), and the potential misuse of LLM-generated findings. Pragmatic concerns with data quality, fairness, and equity of access to the powerful AI systems will be substantial.

In deciding whether to use LLMs to approximate human behavior, researchers must first validate language-mediated (latent) constructs (*2*). They can treat LLM-generated responses as a "sample" of nonhuman participants and systematically vary prompts, akin to presenting random stimuli in traditional experiments. A crucial consideration in using LLMs for research is the trade-off between external and internal validity. Future LLMs, trained on diverse cultural content, will offer greater external validity by simulating human-like responses and generalizing to real-world scenarios. However, their opaque nature will limit their internal validity. Conversely, laboratory-grown natural-language processing models, built on smaller controlled datasets, will provide stronger internal validity at the expense of reduced reliability and generalizability because the limited training data may hinder their ability to perform consistently and broadly across different contexts. Researchers should carefully choose between these approaches according to their priorities.

"...large language models rely on 'shadows' of human experiences described in cultural products."

Researchers must also consider the context of their study. High-risk situations that involve violence or situations that are plainly infeasible with large numbers of human participants may be more suitable for LLMs. For example, LLMs might be used to explore human dynamics of space travel, or create predator and victim prototypes for studies of online sexual predators, an ethically fraught realm because of the potential trauma to human participants.

As AI reshapes the landscape of social science (*14*), researchers will diversify their expertise, embracing new roles such as model bias hunters, AI-data validators, or human-AI interactionist. In this context, maintaining conceptual clarity (*2*), understanding foundations of measurement (*2*), and adhering to ethically grounded practical wisdom (*15*) for selecting an AI-assisted design that fits one's research question will be essential. With the democratization of AI-assisted data collection, the importance of early-stage social science training and supporting quantitative methods (such as computation and statistics) is crucial, calling for revision of social science education programs.

Just as the prisoners in the allegory of Plato's Cave observe shadows on a wall and believe them to represent reality, LLMs rely on "shadows" of human experiences described in cultural products. These shadows offer a limited view of the true nature of the phenomena they represent because folk psychology (*2*) captured in cultural products may not always reflect the mechanisms that govern human behavior—a limitation essential for social scientists to acknowledge. Examining the limitations and biases of LLMs also puts a mirror to common practices in many fields, be it bias in representation, sampling methods, or methodological individualism (*2*).

Despite these obstacles, LLMs allow social scientists to break from traditional research methods and approach their work in innovative ways. LLM models will likely vitiate online crowdworking platforms, which are the dominant source of human participant data in many social science fields, for the simple reasons of on-par performance of simple tasks, and because open-ended responses from LLM-guided bots will become indistinguishable from human participants, calling for new methods for human data verification. Social scientists must be prepared to adapt to the uncertainty (*15*) that comes with evolving technology while being mindful of the limitations of ongoing research practices. Only by maintaining transparency and replicability (*2*) can we ensure that AI-assisted social science research truly contributes to our understanding of human experience. ■

REFERENCES AND NOTES

1. S. Bubeck *et al*., *arXiv*:2303.12712 [cs.CL] (2023).
2. Extended documentation of LLMs abilities, ethical challenges, and methodological concerns, along with foundational social science principles, is available on Open Science Framework (https://osf.io/h4e2a).
3. L. P. Argyle *et al*., *Polit. Anal*. 10.1017/pan.2023.2 (2023).
4. J. S. Park *et al*., *arXiv*:2304.03442 [cs.HC] (2023).
5. P. E. Tetlock, C. B. McGuire, G. Mitchell, *Annu. Rev. Psychol*. **42**, 239 (1991).
6. H. Shirado, N. A. Christakis, *Nature* **545**, 370 (2017).
7. G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, D. G. Rand, *Psychol. Sci*. **31**, 770 (2020).
8. M. Galesic, H. Olsson, J. Dalege, T. van der Does, D. L. Stein, *J. R. Soc. Interface* **18**, 20200857 (2021).
9. P. Antosz, S. Bharwani, M. Borit, B. Edmonds, *Int. J. Soc. Res. Methodol*. **25**, 511 (2022).
10. A. Abid, M. Farooqi, J. Zou, *Nat. Mach. Intell*. **3**, 461 (2021).
11. S. Fazelpour, D. Danks, *Philos. Comp.* **16**, e12760 (2021).
12. Y. Bai *et al*., *arXiv*:2212.08073 [cs.CL] (2022).
13. L. Weidinger *et al*., in *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2022), pp. 214–229.
14. J. C. Peterson, D. D. Bourgin, M. Agrawal, D. Reichman, T. L. Griffiths, *Science* **372**, 1209 (2021).
15. I. Grossmann *et al*., *Psychol. Inq*. **31**, 103 (2020).

ACKNOWLEDGMENTS

The authors thank T. Charlesworth, R. Saxe, and S. Fazelpour for their feedback on earlier versions of the draft. This work was funded by Social Sciences and Humanities Research Council of Canada Connection grant 611-2020-0190 (to I.G.), Social Sciences and Humanities Research Council of Canada Insight grant 435-2014-0685 (to I.G.), and John Templeton Foundation grant no. 62260 (to I.G. and P.E.T.).

10.1126/science.adi1778