Exploiting LLMs to Better Understand Assumptions in Social Science Methodologies

Nedah Nemati, Presidential Scholar in Society and Neuroscience

Center for Science and Society, Columbia University

Introduction

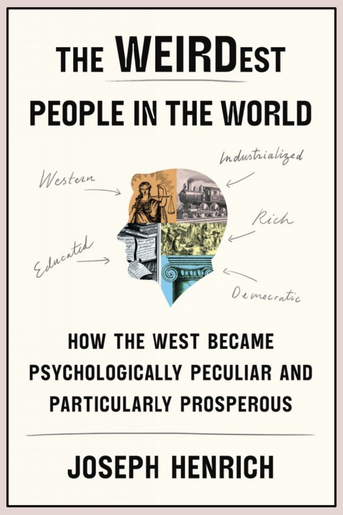
Measuring 'Man': Using social scientific techniques to measure and analyze humans has long been facilitated and limited by methods of idealization and abstraction

Homo Economicus | *Homo Psychologicus* | *Homo Silicus*

Prisoner's Dilemma

Outcome	Y Cooperates	Y Defects
X Cooperates	(1,1)	(5,0)
X Defects	(0,5)	(3,3)

Tragedy of the Commons



Silicon Sampling: Conditioning a large language model (LLM) to simulate real people and polling these samples for analysis.

- In this case, we are exploring studies that poll the 'silicon' person to gauge political orientations, opinions, and voting behaviors. However, 'silicon sampling' can be used for a wide variety of tasks.

Idealization and Abstraction in the Social Sciences

Argyle et al. (2023) conditioned their model on thousands of socio-demographic backstories from U.S. American survey responses (3).

The authors argue for the model's *algorithmic fidelity*: "the degree to which the complex patterns of relationships between ideas, attitudes, and socio-cultural contexts within a model accurately mirror those within a range of human subpopulations" (4):

1. Social Turing Test
2. Backward Continuity
3. Forward Continuity
4. Pattern Correspondence

Using a battery of tests, the authors show that the model meets algorithmic fidelity, concluding that it offers "general-purpose windows into human thinking" by revealing the "many various patterns of associations between ideas, attitudes, and contexts present among humans" (2, 15).

Idealization and Abstraction as Acts of Imagination

Question: How do social scientists' tools orient to permit certain imaginations over others?

- In the case of silicon sampling, what methodological constraints (including assumptions) are exacerbated by the use of LLMs?

Ethical Challenges to Silicon Sampling

Challenge 1

A tradeoff exists between the accuracy of a silicon sample and training LLMs on harmful data.

Scenario A:

To maintain ethical standards, the LLM is trained on data deemed 'safe' (i.e., non-racist, non-sexist, etc.)

Scenario B:

LLM is trained on a wide range of data (including harmful data), but because of efforts to reduce harm, model output is censored or even further "morally self corrected" (Ganguli et al., 2023).

In both scenarios, the LLM fails to represent the full range of human 'types' for social research.

Challenge 2

Similar to other AI, silicon sampling can be used for harmful persuasion tactics.

Argyle et al. briefly note the potential for mis-information, fraud, and manipulation.

Political and Legal Action

Figure 1. More pressing are issues related to positive and negative influence operations (Goldstein et al., 2023). Pictured here, the President of Slovenia, Nataša Pirc Musar, discusses issues of persuasion tactics as they relate to LLMs (Columbia University, March 23, 2023).

Challenge 3

The 'success' of silicon sampling can both conflate the various goals of social scientific questions and promote realist commitments to the social phenomena under question.

Silicon samplers claim that the success of the models captures something deep about human capacities. If we grant the success of their experiments, Argyle et al. posit that their model provides access to a deeper understanding of political beliefs and human traits.

Such claims not only risk reifying realist assumptions about human behaviors and cognition, but associated assumptions can result in 'looping effects' regarding the phenomena under investigation (Hacking, 2002). See Figure 2 below for a historical example.

Unanticipated Payoffs and Future Explorations

- Interrogating silicon sampling methodology, as well as exploring the ethical issues generated by these procedures, reveals how LLM generated silicon samples can be used as a tool for philosophers of science who are reflecting on the limitations of social science methodology
  - ➢ Challenges 1 and 3 make explicit the pressures that social scientists contend with when choosing who to sample. Challenge 3 also draws attention to the issue of conflating researcher goals; in the case of Argyle et al., significant differences arise if the social scientific question posed is for the purpose of an intervention or for characterizing demographic groups (this is a limitation of only using sampling for polling purposes)
- Taking these challenges seriously, in turn: 1) raises questions about the anthropocentric approaches taken in social science research (see Epstein, 2015), and 2) de-emphasizes the importance of the social scientists' toolkits for driving actionable results. An open question thus emerges about *what else* is involved in such experimentation
- Ironically, LLMs are showing that if silicon sampling is 'successful', it is not because of the tools themselves, but rather because of other factors and decisions driven by the researchers. Abstracting away from this can miss the point of the social project
- A promising vein of future studies would be a comparative historical and philosophical analysis of computational model systems that have been used as tools for scientific research

Biologically Inspired Neural Nets

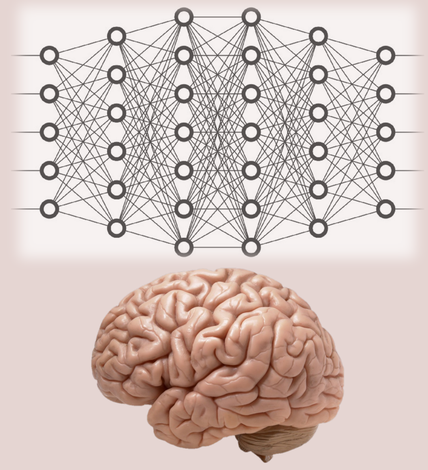


Figure 2. This figure provides an example of 'Challenge 3'. The history of neural networks and their biological fidelity demonstrates how overextending claims from computational models creates realist commitments and 'backwards idealization' (i.e., idealizing from the brain to idealizing from the model back to the brain).

References

Argyle, Lisa P., et al. "Out of one, many: Using language models to simulate human samples." *Political Analysis* (2022): 1-15.
Epstein, Brian. *The ant trap: Rebuilding the foundations of the social sciences*. Oxford Studies in Philosophy, (2015).
Ganguli, Deep, et al. "The Capacity for Moral Self-Correction in Large Language Models." *arXiv preprint arXiv:2302.07459* (2023).
Goldstein, Josh A., et al. "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations." *arXiv preprint arXiv:2301.04246* (2023).
Hacking, Ian. "Making Up People." *Historical Ontology*, Harvard University Press, 2002, pp. 99–114. *JSTOR*.

Acknowledgements

This project began in fruitful conversations with Will Conner. I am also grateful to Mahi Hardalupas and Os Keyes for conceptual feedback.