Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025)

LLM-based Simulations of Human Behavior in Psychological Research

Santiago Flórez Sánchez
Universidad de los Andes
s.florezs2@uniandes.edu.co

Abstract

What does it mean for LLMs to replace human participants in psychological research? My analysis of this question is structured around two central philosophical problems: scientific representation and epistemic opacity. By examining how these issues shape trustful and distrustful stances toward using LLMs as models of the human mind, I describe tendencies in the scientific literature and their relation to emerging interpretability and elicitation techniques. In this regard, my primary contributions are, first, a philosophical framework for understanding the conceptual tensions that shape the debate, and second, a taxonomy that maps stances in empirical literature to their corresponding methodological innovations. I show that both trustful and distrustful positions, despite their disagreements, foster the methodological innovations necessary for building a more robust epistemological foundation for LLM-based simulations. Accordingly, empirical research stances must be responsive to the pressures and constraints implied by their underlying philosophical intuitions. This means, for instance, that trustful stances should explore protocols leveraging fine-tuning and prompt design to evaluate correspondence and consistency in more complex behavioral patterns—thereby working around model opacity—while distrustful stances should further develop parallels at the algorithmic and implementational levels between LLMs and the human mind through XAI techniques and computational cognitive science—to probe the representational relationship.

A recent debate has centered on whether large language models (LLMs) can replace human participants in psychological research. Optimists argue that LLMs offer cost-effective and rapid behavioral simulations that reflect the psychology embedded in their training data. Skeptics, however, question the representational relationship between computational language models and the human mind. Even if LLMs can mimic linguistic behavior, this does not entail that they are scientific representations of human psychology; they may resemble aliens or zombies.

In this paper, I argue that positions in the debate over replacing human participants are shaped by underlying assumptions about two philosophical issues: scientific representation and epistemic opacity. I also highlight the challenges raised by these perspectives, as well as the techniques employed to address them. Then, I map these philosophical insights onto the conditions proposed in the scientific literature for a simulation to be considered a valid representation of human behavior.

While LLMs can be valuable exploratory tools in psychology, they are constrained by the epistemic opacity inherent in their technical design. Accordingly, they should be understood as heuristic models of human behavior that require complementary methods of analysis and enhancement. However, this debate is fostering the development of novel supplementary methodologies, which will nourish the scientific ecosystem and advance the epistemological foundations of LLM-based simulations. Therefore, it is crucial to situate empirical research stances within their philosophical commitments and methodological context, in order to recognize productive lines of inquiry.

1. Philosophical Problems

The debate over replacing human participants with LLMs touches upon two distinct yet interconnected issues in the philosophy of science: the problem of scientific representation and the problem of epistemic opacity.

1.1 Scientific Representation

In the philosophy of science, a model is a scientific tool used to investigate a much broader or more complex phenomenon. But what does it mean for a model to represent a target phenomenon? The foundation of scientific representation lies in *surrogative reasoning*—the possibility that a scientist may gain knowledge about the target through the analysis of

the model (Frigg and Nguyen 2020). Identifying the conditions that make surrogative reasoning possible is central to our debate, since LLM-based simulations aim to yield knowledge about human psychology as if they were actual participant data. In other words, justifying the representational nature of the model requires demonstrating that human psychology can be studied through LLM-based simulations. In this sense, the idea of replacing human participants hinges on the question of whether LLMs represent human behavior well enough to function as scientific models of the mind.

Two common views on the problem of scientific representation can be distinguished. *Naturalist* positions hold that representation is a factual or objective relation that depends directly on the properties of both the model and the target phenomenon, rather than on the agents' scientific intentions (Suárez 2003). Within this tradition, proposals based on similarity or isomorphism argue that a model represents its target by sharing properties with it, or more specifically, by mirroring its abstract structure (Frigg and Nguyen 2020). For instance, a scale model of a bridge represents the bridge insofar as it resembles it in the relevant features. Likewise, validating a representational relation between an LLM and the human mind requires justifying their connection—be it resemblance, analogy, or any other proposed relation—and this justification must refer to the properties of the systems in question.

In contrast, *non-naturalist* positions conceive of representation as an artificial relation that includes the agent, the context, and the scientific practice, such that what constitutes representation goes beyond the model's properties. A central view within this tradition is inferentialism, which proposes that scientific intention and the model's inferential capacity are necessary conditions for representation (Suárez 2004). Another noteworthy view is artifactualism, which holds that the process of model construction underpins the justification for its representational character. Emphasizing the scientific practice broadens the range of possible representational means and focus the concerns on the corresponding methodological questions for each case. These approaches seek to dissolve the problem by no longer requiring an independent foundation for the model's representational status beyond its actual use. From this perspective, the use of LLMs as models of the human mind could be justified by their crafting process and the inferences they allow, depending on the methodological details of each case.

1.2 Epistemic Opacity

The concept of *epistemic opacity* was originally introduced in the early 2000s in the context of computational simulations, but it remains relevant—and arguably even more pressing—in the case of deep learning (DL) models. A process is epistemically opaque when its epistemically relevant elements are unknown (Humphreys 2009 4). "The problem arises from the lack of an explicit algorithm linking the initial inputs with the final outputs, together with the inscrutability of the hidden units that are initially trained" (Humphreys 2004 149). Since DL models—including LLMs—operate and learn through large-scale, autonomous, and stochastic computations that elude human understanding (perhaps even by their very nature), they are considered opaque and are often compared to black boxes. Not only are the step-by-step processes that generate outcomes unclear; in DL models, the internal representations instantiated in the neural network parameters—i.e., what the model "learned" during training—are entirely unknown (Boge 2022). In this sense, the issue diverges from the conditions of surrogative reasoning but highlights the inherent complexity of these models, and it often fuels the critique that black-box systems lack scientific utility.

One proposed response to this problem is *reliabilism*. According to this view, an opaque computational simulation may be justified if it is reliable, meaning that the simulation process has an objectively high likelihood of functioning properly, based on prior observations of its performance (Durán and Formanek 2018). What counts as proper functioning and what the sources of reliability are may vary depending on the methodological context. Still, this idea directly suggests predictive accuracy as a criterion for the representational legitimacy of an opaque computational model (Lenhard 2009). Even if the internal mechanisms remain unknown, one may rely on the model's representational value based on its results. Therefore, the justification for using LLMs as replacements for human participants in psychological research would depend on how effectively they replicate human data.

Conversely, a *non-reliabilist* position would argue that the justification for an opaque computational simulation cannot rest solely on the alignment of its outputs with real data. It must also appeal to the model's processing mechanisms. This view may be motivated by the concern that one can arrive at the correct results through flawed procedures—even though, again, what counts as "flawed" is debatable. This focus on internal processing makes the non-reliabilist stance more sensitive to the problems posed by epistemic opacity, as it is unwilling to accept a model without also examining how it works. In other words, for this position, epistemic opacity is something to be tackled—not sidestepped.

***

In summary, the first problem questions what kind of relation scientific representation is, and gives rise to a distinction between *natural* relations, based solely on the properties of the model and the target, and *non-natural* ones, which depend on contextualized scientific practice. The second

problem addresses the grounds for justifying the representational value of an opaque model and leads to a distinction between *reliabilist* justifications, based solely on results, and *non-reliabilist* ones, which appeal to the model's internal processing.

2. Mapping the Debate Space

Broadly speaking, in the debate on replacing human participants, one can distinguish between trustful and distrustful stances. Distrustful positions are less willing to accept LLM-based simulations without epistemic guarantees, while trustful positions are more open to endorsing a representational relationship between language models and the human mind. This description functions as a map of the debate space in the following sense: the earlier philosophical problems define axes of discussion, the positions outlined below are demarcated by their representational conditions—that is, by how they respond to those philosophical problems—and the typical stances illustrate concrete tendencies within the debate.

2.1. Distrustful Criteria

Distrustful stances tend to endorse stricter conditions for scientific representation. The stricter these conditions, the harder it is to accept that a computational model can replace a human participant. Such criteria may be stricter in two ways: by attempting to warrant a natural relationship between the LLM and the mind, or by requiring a justification of the model that appeals to its internal processing. Justifying a natural relationship is, in principle, more challenging because it requires engaging with the properties of both the model and the target, and must therefore be grounded in theoretical frameworks and empirical evidence—something particularly difficult given our limited understanding of, and lack of consensus about, the mind. Appealing to internal processing is likewise difficult, as it requires overcoming the epistemic opacity inherent to DL models.

A typical distrustful position expects the representational character of an LLM to be justified through its internal processing, since this could establish a relationship with the mind in terms of the model's properties. In other words, the driving force behind skepticism is that epistemic opacity hinders model validation. There are different positions on what would validate the model in terms of its properties. A weaker naturalism might accept a less detailed interpretation of the algorithm, while a stronger naturalism might argue that an LLM cannot scientifically represent the human mind unless its internal processing also simulates biological processing. In turn, even if a distrustful stance does not require the representational relationship to be fully grounded in its properties (a non-naturalist position), it may still be dissatisfied with a justification that completely disregards the model's internal processing.

From a distrustful perspective, addressing epistemic opacity requires some form of understanding of the algorithm computed by the language model. The underlying intuition is that an LLM is a good representation—at least partially—insofar as its algorithm corresponds to human mental processes. These skeptical concerns motivate technical interpretations based on architecture and training, efforts to identify parallels between the human mind and LLMs, and the use of eXplainable AI (XAI) techniques to provide evidence for interpretation. For example, if LLMs are conceived as reproducing linguistic and semantic patterns from their training data, one could attempt to relate that mechanism to human cognitive processes of linguistic comprehension or production; or given that ANNs are inspired by biological neural networks, one could argue that this algorithmic design simulates the human processing. However, these generic interpretations neither explains why reproducing linguistic patterns or fundamental cognitive architecture should capture causal relationships in human psychology nor accounts for the model's internal representations. More fine-grained interpretative hypotheses would instead be specific to concrete use cases and more demanding to empirically validate.

2.2. Trustful Criteria

On the other hand, trustful stances adopt more flexible conditions of representation—for instance, methodological precautions and their related predictive accuracy. These conditions may be more flexible either because they consider the scientific context in which the LLM is constructed and used as a model of the mind, or because they accept justifications based exclusively on the model's outputs. Accounting for the scientific practice in which the model is embedded facilitates its justification by shifting the problem toward context-specific methodological or logical issues, thereby avoiding theoretical presuppositions. In turn, appealing exclusively to the model's results facilitates justification by evading the limitations posed by epistemic opacity.

A typical trustful position holds that its justification need not go beyond the model's results, since representational character does not depend exclusively on the model's internal properties but rather on its inferential (or predictive) capacity. Thus, if LLM-based simulations match psychologically relevant patterns of human behavior, this is considered evidence that they represent the human mind—at least within the concrete circumstances of the simulation. That said, there are various interpretations of what counts as matching or enabling inference: from the Turing test, to opinion alignment, to background coherence or behavioral pattern reproduction. Alternatively, a trustful variant might

even accept that a natural relationship is necessary for scientific representation, while still taking predictive accuracy as an indication of such a connection with the target.

Underlying this view is the idea that predictive success implies that the algorithm instantiated by the model captures causal-explanatory relationships of the phenomenon (Pietsch 2016; Lenhard 2009). This idea has led to several consequences. First, the race for higher predictive accuracy has driven the development of novel elicitation techniques. However, these techniques have shown that LLM-based simulations are not always as cost-effective and rapid as initially expected. Second, the lack of understanding of implementation details (i.e., the step-by-step of opaque computations) does not necessarily preclude grasping the structure of the computed algorithm (Sullivan 2022). Third, this view raises important questions about whether and why detailed algorithmic interpretation is needed to justify a model's representational value, and thus how threatening epistemic opacity really is. Still, one might ask: why is it compelling to claim that nothing can be explained by an unexplained model?

3. Scientific Literature

To illustrate, consider how the stances characterized in the previous section map onto the scientific literature addressing our debate. Current research aims to interpret the inner representations of LLMs and to assess the reliability of their simulations. In general, experiments tend to compare LLMs and elicitation techniques, while attempting to correlate simulations with actual human behavior in a given task and, in some cases, employ interpretability XAI techniques (frequently post-hoc, attribution methods).

3.1. Distrustful Literature

The relationship between computational language models and human cognitive processing has sparked debates at both the algorithmic and implementational levels. To tackle the epistemic opacity of LLMs (i.e., to interpret their internal processing) researchers often rely on available information about model architecture and training, as well as XAI techniques. These interpretations vary in both generality—from addressing the processing of language itself in comprehension or production tasks, to focusing on specific cognitive capacities or patterns (e.g., reasoning or priming)—and methodology—from applying psychological tests or experiments to LLMs to compare known human effects or traits, to postulating computational analogies or correlating processing patterns with brain measurements.

Studies aiming to establish algorithmic-level relationships between LLMs and the human mind argue for interpretations of the algorithm computed by the LLM and compare it with what is known about human cognition. Such discussions may compare processing in terms of the conceptual representations involved (Yetman 2025; Xu et al. 2025), its computational principles, functional properties, or features of cognitive architecture (Mischler et al. 2024; Niu et al. 2024; Caucheteux and King 2021, 2022; Goldstein et al. 2022). These studies typically include empirical evidence, conclude with a partial parallel, and are often accompanied by technical recommendations to refine the similarity.

Other studies aim to demonstrate implementation-level relationships between the computations carried out by these cognitive systems. A popular methodology consists in measuring the alignment between LLM embeddings and human brain activity in regions relevant to the task at hand. Using this approach, researchers have drawn parallels between activity patterns in certain layers of LLMs (represented by their internal embeddings) and neural activity patterns in specific regions and layers of the cortex—in relation to language comprehension and production tasks across different modalities, as well as more specific capacities such as Theory of Mind (ToM) (Ren et al. 2025; Caucheteux and King 2021, 2022; Mischler et al. 2024; Rahimi, Yaghoobzadeh, and Daliri 2025; Cai et al. 2025; Goldstein et al. 2025; Jamali, Williams, and Cai, 2023).

Clearly, XAI plays a significant role in this body of literature on similarity between LLMs and the human mind and brain. Even though these interpretability techniques have been strongly criticized, show substantial limitations, and remain under development (Sullivan 2024; Longo et al. 2024), they have garnered attention for their potential as methodologies for cognitive science and neuroscience (Zednik and Boelsen 2022; Cai et al. 2025; Caucheteux and King 2021, 2022; Goldstein et al. 2022; Mischler et al. 2024; Niu et al. 2024; Rahimi, Yaghoobzadeh, and Daliri 2025; Ren et al. 2025). It is worth noting that, within the debate on taking LLMs as models of the mind, these emerging techniques are motivated by skeptical discomfort with epistemic opacity in DL models and/or by a naturalistic drive to compare their properties with those of the brain.

3.2. Trustful Literature

Trustful stances tend to focus on the model's predictive capacity and the scientific practice in which it is embedded. This brings into play several factors involved in the construction, use, and evaluation of the model. Since LLMs are DL models, LLM-based simulations must be evaluated to verify, for example, that they genuinely extrapolate semantic connections from training data, avoiding memorization or overfitting (Eigenschink et al. 2023). Regarding use and construction, key methodological factors include prompting and fine-tuning techniques. Additional variables involved in LLM use are generation parameters such as temperature, max-length, top-p or top-k sampling, etc. Furthermore, part

of the model's construction involves the curation and justification of training or tuning datasets. This is accompanied by methodological concerns, for example, about the replicability of experiments or the diversity of simulated data (Niszczota, Janczak, and Misiak 2025; Wang, Morgenstern, and Dickerson 2025; Abdurahman et al. 2024a, 2024b; Demszky et al. 2024; Santurkar et al. 2023; Park, Schoenegger, and Zhu 2023).

Still, to evaluate predictive accuracy—and hence inferential capacity—the primary representational criteria are correspondence and consistency, and the main evaluation methods are comparison with real data and human evaluation. Several studies propose specific lists of criteria, but they can generally be reduced to these two categories.

Correspondence refers to the emulation of human behavioral patterns, which may include opinions, decisions, performance, or cognitive and behavioral biases (Strachan et al. 2024; Rossi, Harrison, and Shklovski 2024; Wang et al. 2024; Aher, Arriaga, and Kalai 2023; Guo 2023; Binz and Schulz 2023a, 2023b; Santurkar et al. 2023; Loconte et al. 2023; Sap et al. 2022). Consistency refers to the coherence of behavior over the course of an interaction (Suzuki and Arita 2024; Sreedhar and Chilton 2024; Jiang et al. 2024; Horton 2023; Xiao et al. 2023; Simmons and Hare 2023; Argyle et al. 2022). This criterion encompasses several dimensions: a model is a) robust if its behavior maintains a consistent identity (Xiao et al. 2023), b) sensitive if it adapts to the interaction (Simmons and Hare 2023), and c) coherent (or continuous, realistic, or simply consistent) if it aligns with the simulated profile (Eigenschink et al. 2023; Simmons and Hare 2023; Xiao et al. 2023; Argyle et al. 2022).

For instance, the Turing test, via human evaluation, aims to determine whether a model behaves like a human (correspondence) over the course of a conversation (consistency). Likewise, researchers directly compare human responses to simulated data from the same test or survey, or analyze aspects of the generated text (e.g., vocabulary, connotation, tone), or test predictors of whether a human would classify a post as belonging to a given demographic group, or assess the likelihood of a given decision given a specific profile.

As with criteria and methodology, results have varied. Some studies have shown high correlation (≥ 0.9) between simulated data and average human results in experiments (Dillon et al. 2023; Aher, Arriaga, and Kalai 2023; Argyle et al. 2022; Binz and Schulz 2023a; Hewitt et al. 2024). In contrast, others have reported inconsistencies not only in opinion alignment but also in capabilities and other behavioral patterns (Santurkar et al. 2023; Gao et al. 2024; Binz and Schulz 2023b; Loconte et al. 2023; Sap et al. 2022; Tjuatja et al. 2024; Salecha et al. 2024; Park, Schoenegger, and Zhu 2023). This latter evidence points, at the very least, to limitations concerning the appropriate use cases for LLM-based simulations.

Ultimately, the most significant advances in this literature concern the ongoing development of elicitation techniques, including prompting and fine-tuning. Prompting techniques are a core element of LLM usage. In addition to being correlated with predictive accuracy, they underpin the simulation of social interactions through multi-agent design (Sreedhar and Chilton 2024; Simmons and Hare 2023; Park et al. 2023, 2022; Argyle et al. 2022), and even enable the extraction of additional information from simulations (e.g., justifications or chains of thought) (see Xu et al. 2024; Guo 2023). Similarly, fine-tuning techniques have consistently proven useful for enhancing LLM performance in specific domains and, thus, predictive accuracy in research contexts (Lu, Luu, and Buehler, 2025; Liu et al. 2025; Zhang et al. 2025; Abdurahman et al. 2024a; Gao et al. 2024; Binz and Schulz 2023a; Horton 2023). Nowadays, for example, researchers can tune a commercial LLM at a relatively low cost with data specific to the behavior being modeled. Within the debate on taking LLMs as models of the mind, these emerging techniques are motivated by an interest in refining the symmetry between the simulation and human behavior.

4. Methodological Implications

Both trustful and distrustful stances should remain aware of the constraints imposed by their respective philosophical intuitions about scientific representation and epistemic opacity. These controversies drive innovation in interpretability and elicitation techniques. In turn, empirical research must also stay informed about methodological innovations in the field, as this is essential for contributing effectively to the debate.

For instance, distrustful stances should be aware of the extensive effort required to empirically justify a natural relationship between cognitive systems that are still poorly understood. While efforts in this direction are promising, they remain incipient and depend on significant advances in cognitive science and neuroscience, making such a requirement unfeasible in the short term. As such, it would be overly rigid to reject all scientific uses of LLM-based simulations because of the absence of sufficient epistemic guarantees. Instead, in addition to continuing the comparative analysis between the internal processes of LLMs and the human mind/brain, researchers should integrate XAI techniques into simulation protocols that use the latest advances in fine-tuning and prompt design.

Likewise, trustful stances should recognize the unease generated by unexplainable models and the domain-specific justification of the representational character of LLMs, as well as the technical difficulties involved in tuning and manipulating LLMs. On the one hand, it is understandable that

simple correlations between model outputs and human opinions may not be compelling enough. Hence, it is advisable to refine experimental protocols to demonstrate the predictive accuracy of simulations in relation to a broad range of behavioral features, especially those that are harder to mimic. On the other hand, the epistemic foundation for this methodology is piecemeal, depending not only on how the model is deployed, but also on the specific subfield of psychology and the mental trait or process being modeled. By contrast, it would be careless to assume that an off-the-shelf commercial LLM can yield insight into human psychology without any control or knowledge regarding its architecture, training, tuning, or underlying data. Similarly, it would be careless to extrapolate predictive success in one mental domain to others without further scrutiny.

Conclusion

To recap, in the debate on replacing human participants with LLMs, trustful and distrustful positions are structured around two central philosophical issues: scientific representation and epistemic opacity. The first concerns the conditions under which surrogative reasoning from an LLM is warranted, while the second underscores the model's inherent complexity that hinders its validation. These axes have been used to map the debate space, describing the motivations behind stances and tendencies in the discussion. It is typical, for instance, to find perspectives that either appeal to the model's internal processing to probe its link with the mind or brain, or emphasize its predictive capacity and focus on methodological aspects of model deployment. Accordingly, trustful stances tend to prioritize empirical evaluations of simulation performance and elicitation techniques, whereas distrustful stances demand more transparent internal explanations (often through interpretability methods).

Both lines of research should be interconnected, as they address interdependent questions and their methodological innovations can be mutually reinforcing. A key recommendation is the widespread integration of interpretability methodologies—including, but not limited to, post-hoc attribution XAI techniques—when testing elicitation methods and simulations more broadly. Further recommendations include considering more sophisticated representation criteria that measure various aspects of the simulated linguistic behavior and "cognitive process"; proposing more challenging mental capacities or phenomena to simulate; and verifying the correspondence of simulations on a case-by-case basis for each field of psychology. Other issues may also be addressed in future research, such as the replicability of simulated data and the generalization of models to new data. However, none of this suggests that language models will eliminate the need to collect human data. On the contrary, training and elicitation techniques reaffirm the value of (up-to-date and well-curated) data.

The controversies and concerns currently fueling the debate can be better understood by identifying deeper philosophical motivations and considering emerging computational techniques. Since contemporary research applies diverse criteria to evaluate the representational relationship between LLMs and the mind, the aforementioned philosophical discussions serve as a framework for understanding the conflicting positions as differing in their requirements for representation. Furthermore, interpretability and elicitation methods play a key role in the methodological grounding and development of LLM-based simulations, as they push back against the empirical limitations imposed by epistemic opacity. Therefore, contributing to research on modeling the mind with LLMs also requires being informed about the technical details of these models and about XAI, prompting, and fine-tuning methods.

In short, LLMs are valuable exploratory tools for psychological research, but they need to be supported by interpretability and elicitation methods. This debate concerns deep philosophical questions and intuitions, and provides fertile ground for the development of novel supplementary methodologies. As part of the effort to build an epistemic foundation for LLM-based simulations of human behavior, it is necessary to identify the underlying assumptions and commitments of the research, as well as the broader challenges they entail. This article seeks to contribute to that effort.

Acknowledgments

I thank Professor William Jiménez for his mentorship in the Cognition and Learning Laboratory, and Professor Andrés Páez for his course "Models, Simulations, and Artificial Intelligence."

References

Abdurahman, S.; Atari, M.; Karimi-Malekabadi, F.; Xue, M. J.; Trager, J.; Park, P. S.; Golazizian, P.; Omrani, A.; Dehghani, M. 2024a. Perils and Opportunities in Using Large Language Models in Psychological Research. *PNAS Nexus* 3(7). doi.org/10.1093/pnasnexus/pgae245

Abdurahman, S.; Ziabari, A. S.; Moore, A.; Bartels, D.; and Dehghani, M. 2024b. A Primer for Evaluating Large Language Models in Social Science Research. PsyArXiv. doi.org/10.31234/osf.io/ag7hy_v2

Aher, S.; Arriaga, R. I.; and Kalai, A. T. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. arXiv. doi.org/10.48550/arXiv.2208.10264

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31(3): 337–351. doi.org/10.1017/pan.2023.2

Binz, M., and Schulz, E. 2023a. Turning Large Language Models into Cognitive Models. arXiv. doi.org/10.48550/arXiv.2306.03917

Binz, M., and Schulz, E. 2023b. Using Cognitive Psychology to Understand GPT-3. *PNAS* 120(6). doi.org/10.1073/pnas.2218523120

Boge, F. J. 2022. Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines* 32(1): 43-75. doi.org/10.1007/s11023-021-09569-4

Cai, J.; Hadjinicolaou, A. E.; Paulk, A. C.; Soper, D. J.; Xia, T.; Wang, A. F.; Rolston, J. D.; Richardson, R. M.; Williams, Z. M.; and Cash, S. S. 2025. Natural Language Processing Models Reveal Neural Dynamics of Human Conversation. *Nature communications* 16(1): 3376. doi.org/10.1038/s41467-025-58620-w

Caucheteux, C., and King, J. R. 2021. Language Processing in Brains and Deep Neural Networks: Computational Convergence and Its Limits. bioRXiv. doi.org/10.1101/2020.07.03.186288

Caucheteux, C., and King, J. R. 2022. Brains and Algorithms Partially Converge in Natural Language Processing. *Communications Biology* 5: 134. doi.org/10.1038/s42003-022-03036-1

Demszky, D.; Yang, D.; Yeager, D. S.; Bryan, C. J.; Clapper, M.; Chandhok, S.; Eichstaedt, J. C.; Hecht, C.; Jamieson, J.; Johnson, M.; Jones, M.; Krettek-Cobb, D.; Lai, L.; JonesMitchell, N.; Ong, D. C.; Dweck, C. S.; Gross, J. J.; and Pennebaker, J. W. (2024). Using Large Language Models in Psychology. *Nature Reviews Psychology* 2: 688-701. doi.org/10.1038/s44159-023-00241-5

Dillion, D.; Tandon, N.; Gu, Y.; and Gray, K. 2023. Can AI Language Models Replace Human Participants? *Trends in Cognitive Sciences* 27(9): 690-702. doi.org/10.1016/j.tics.2023.04.008

Durán, J. M., and Formanek, N. 2018. Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism. *Minds and Machines* 28: 645–666. doi.org/10.1007/s11023-018-9481-6

Eigenschink, P.; Reutterer, T.; Vamosi, S.; Vamosi, R.; Sun, C.; and Kalcher, K. 2023. Deep Generative Models for Synthetic Data: A Survey. *IEEE Access* 11: 47304-47320. doi.org/10.1109/ACCESS.2023.3275134

Frigg, R., and Nguyen, J. 2020. Modelling Nature: An Opinionated Introduction to Scientific Representation. Switzerland: Springer. doi.org/10.1007/978-3-030-45153-0

Gao, Y.; Lee, D.; Burtch, G.; and Fazelpour, S. 2024. Take Caution in Using LLMs as Human Surrogates: Scylla ex machina. arXiv. doi.org/10.48550/arXiv.2410.19599

Goldstein, A.; Zada, Z.; Buchnik, E.; Schain, M.; Price, A.; Aubrey, B.; Nastase, S. A.; Feder, A.; Emanuel, D.; Cohen, A.; Jansen, A.; Gazula, H.; Choe, G.; Rao, A.; Kim, C.; Casto, C.; Fanda, L.; Doyle, W.; Friedman, D.; Dugan, P.; Melloni, L.; Reichart, R.; Devore, S.; Flinker, A.; Hasenfratz, L.; Levy, O.; Hassidim, A.; Brenner, M.; Matias, Y.; Norman, K. A.; Devinsky, O.; and Hasson, U. 2022. Shared Computational Principles for Language Processing in Humans and Deep Language Models. *Nature Neuroscience* 25: 369–380. doi.org/10.1038/s41593-022-01026-4

Goldstein, A.; Wang, H.; Niekerken, L.; Schain, M.; Zada, Z.; Aubrey, B.; Sheffer, T.; Nastase, S. A.; Gazula, H.; Singh, A.; Rao, A.; Choe, G.; Kim, C.; Doyle, W.; Friedman, D.; Devore, S.; Dugan, P.; Hassidim, A.; Brenner, M.; Matias, Y.; Devinsky, O.; Flinker, A.; and Hasson, U. 2025. A Unified Acoustic-to-Speech-to-Language Embedding Space Captures the Neural Basis of Natural Language Processing in Everyday Conversations. *Nature Human Behavior* 9: 1041-1055. doi.org/10.1038/s41562-025-02105-9

Guo, F. 2023. GPT in Game Theory Experiments. arXiv. doi.org/10.48550/arXiv.2305.05516

Guo, Z.; Lai, A.; Thygesen, J. H.; Farrington, J.; Keen, T.; and Li, K. 2024. Large Language Models for Mental Health Applications: Systematic Review. *JMIR Mental Health* 11. doi.org/10.2196/57400

Hewitt, L.; Ashokkumar, A.; Ghezae, I.; and Willer, R. 2024. Predicting Results of Social Science Experiments Using Large Language Models. samim.io/dl/Predicting%20results%20of%20social%20science%20experiments%20using%20large%20language%20models.pdf

Horton, J. J. 2023. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? arXiv. doi.org/ 10.48550/arxiv.2301.07543

Humphreys, P. 2004. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method.* Oxford: Oxford University Press.

Humphreys, P. 2009. The Philosophical Novelty of Computer Simulation Methods. *Synthese* 169: 615-626.

Jamali, M.; Williams, Z. M.; and Cai, J. 2023. Unveiling Theory of Mind in Large Language Models: A Parallel to Single Neurons in the Human Brain. arXiv. doi.org/10.48550/arXiv.2309.01660

Jiang, H.; Zhang, X.; Cao, X.; Breazeal, C.; Roy, D.; and Kabbara, J. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits arXiv. doi.org/10.48550/arXiv.2305.02547

Lenhard, J. 2009. The Great Deluge. Simulation Modeling and Scientific Understanding. In *Scientific Understanding. Philosophical Perspectives*, edited by H. W. de Regt, S. Leonelli, and K. Eigner, 169-186. Pittsburgh: University of Pittsburgh Press.

Liu, B.; Li, X.; Zhang, J.; Wang, J.; He, T.; Hong, S.; Liu, H.; Zhang, S.; Song, K.; Zhu, K.; Cheng, Y.; Wang, S.; Wang, X.; Luo, Y.; Jin, H.; Zhang, P.; Liu, O.; Chen, J.; Zhang, H.; Yu, Z.; Shi, H.; Li, B.; Wu, D.; Teng, F.; Jia, X.; Xu, J.; Xiang, J.; Lin, Y.; Liu, T.; Liu, T.; Su, Y.; Sun, H.; Berseth, G.; Nie, J.; Foster, I.; Ward, L.; Wu, Q.; Gu, Y.; Zhuge, M.; Liang, X.; Tang, X.; Wang, H.; You, J.; Wang, C.; Pei, J.; Yang, Q.; Qi, X.; and Wu, C. 2025. Advances and Challenges in Foundation Agents: From Brain-Inspired Intelligence to Evolutionary, Collaborative, and Safe Systems. arXiv. doi.org/10.48550/arXiv.2504.01990

Loconte, R.; Orrù, G.; Tribastone, M.; Pietrini, P.; and Sartori, G. 2023. Challenging ChatGPT's "Intelligence" with Human Tools: A Neuropsychological Investigation on Prefrontal Functioning of a Large Language Model. doi.org/10.2139/ssrn.4471829

Longo, L.; Brcic, M.; Cabitza, F.; Choi, J.; Confalonieri, R.; Del Ser, J.; Guidotti, R.; Hayashi, Y.; Herrera, F.; Holzinger, A.; Jiang, R.; Khosravi, H.; Lecue, F.; Malgieri, G.; Páez, A.; Samek, W.; Schneider, J.; Speith, T.; and Stumpf, S. 2024. Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions. *Information Fusion* 106. doi.org/10.1016/j.inffus.2024.102301

Lu, W., Luu, R. K. and Buehler, M. J. 2025. Fine-Tuning Large Language Models for Domain Adaptation: Exploration of Training Strategies, Scaling, Model Merging and Synergistic Capabilities. *npj Computational Materials* 11: 84. doi.org/10.1038/s41524-025-01564-y

Mischler, G.; Li, Y. A.; Bickel, S.; Mehta, A. D.; and Mesgarani, N. 2024. Contextual Feature Extraction Hierarchies Converge in Large Language Models and the Brain. *Nature Machine Intelligence* 6: 1467-1477. doi.org/10.1038/s42256-024-00925-4

Niszczota, P.; Janczak, M.; and Misiak, M. 2025. Large Language Models Can Replicate Cross-Cultural Differences in Personality. *Journal of Research in Personality* 115: 104584. doi.org/10.1016/j.jrp.2025.104584

Niu, Q.; Liu, J.; Bi, Z.; Feng, P.; Peng, B.; Chen, K.; Li, M.; Yan, L. K. Q.; Zhang, Y.; Yin, C. H.; Fei, C.; Wang, T.; Wang, Y.; Chen, S.; and Liu, M. 2024. Large Language Models and Cognitive Science: A Comprehensive Review of Similarities, Differences, and Challenges. arXiv. doi.org/10.48550/arXiv.2409.02387

Park, P. S.; Schoenegger, P.; and Zhu, C. 2023. Diminished Diversity-of-Thought in a Standard Large Language Model. arXiv. doi.org/10.48550/arXiv.2302.07267

Park, J. S.; Popowski, L.; Cai, C.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology. New York: Association for Computing Machinery. doi.org/10.1145/3526113.3545616

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv. doi.org/10.48550/arXiv.2304.03442

Pietsch, W. 2016. The causal nature of modeling with big data. *Philosophy and Technology* 29: 137-171. doi.org/10.1007/s13347-015-0202-2

Rahimi, M.; Yaghoobzadeh, Y.; and Daliri, M. R. 2025. Explanations of Large Language Models Explain Language Representations in the Brain. arXiv. doi.org/10.48550/arXiv.2502.14671

Ren, Y.; Jin, R.; Zhang, T.; and Xiong, D. 2025. Do Large Language Models Mirror Cognitive Language Processing? arXiv. doi.org/10.48550/arXiv.2402.18023

Rossi, L.; Harrison, K.; and Shklovski, I. 2024. The Problems of LLM-generated Data in Social Science Research. *Sociologica* 18(2). doi.org/10.6092/issn.1971-8853/19576

Salecha, A.; Ireland, M. E.; Subrahmanya, S.; Sedoc, J.; Ungar, L. H.; and Eichstaedt, J. C. 2024. Large Language Models Show Human-like Social Desirability Biases in Survey Responses. arXiv. doi.org/10.48550/arXiv.2405.06058

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose Opinions Do Language Models Reflect? arXiv. doi.org/10.48550/arXiv.2303.17548

Sap, M.; LeBras, R.; Fried, D.; and Choi, Y. 2022. Neural Theory-of-Mind? On the Limits of Social Intelligence in LLMs. arXiv. doi.org/10.48550/arXiv.2210.13312

Simmons, G., and Hare, C. 2023. Large Language Models as Subpopulation Representative Models: A Review. arXiv. doi.org/10.48550/arXiv.2310.17888

Sreedhar, K., and Chilton, L. 2024. Simulating Human Strategic Behavior: Comparing Single and Multi-Agent LLMs. arXiv. doi.org/10.48550/arXiv.2402.08189

Strachan, J. W. A.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; Manzi, G.; Graziano, M. S. A.; and Becchio, C. 2024. Testing Theory of Mind in Large Language Models and Humans. *Nature Human Behavior* 8: 1285-1295. doi.org/10.1038/s41562-024-01882-z

Suárez, M. 2003. Scientific Representation: Against Similarity and Isomorphism. *International Studies in the Philosophy of Science* 17(3): 225-244. doi.org/10.1080/0269859032000169442

Suárez, M. 2004. An Inferential Conception of Scientific Representation. *Philosophy of Science* 71(5): 767-779. doi.org/10.1086/421415

Sullivan, E. 2022. Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science* 73(1): 109-133.

Sullivan, E. 2024. SIDEs: Separating Idealization from Deceptive 'Explanations' in xAI. In The 2024 ACM Conference on Fairness, Accountability, and Transparency. Rio de Janeiro: ACM. https://doi.org/10.1145/3630106.3658999

Suzuki, R., and Arita, T. 2024. An Evolutionary Model of Personality Traits Related to Cooperative Behavior Using a Large Language Model. *Scientific Report* 14: 5989. doi.org/10.1038/s41598-024-55903-y

Tjuatja, L.; Chen, V.; Wu, S. T.; Talwalkar, A.; and Neubig, G. 2024. Do LLMs Exhibit Human-Like Response Biases? A Case Study in Survey Design. arXiv. doi.org/10.48550/arXiv.2311.04076

Wang, A.; Morgenstern, J.; and Dickerson, J. P. 2024. Large Language Models that Replace Human Participants Can Harmfully Misportray and Flatten Identity Groups. arXiv. doi.org/10.48550/arXiv.2402.01908

Wang, L.; Zhang, J.; Yang, H.; Chen, Z.; Tang, J.; Zhang, Z.; Chen, X.; Lin, Y.; Song, R.; Zhao, W. X.; Xu, J.; Dou, Z.; Wang, J.; and Wen, J.-R. 2023. User Behavior Simulation with Large Language Model Based Agents. arXiv. doi.org/10.48550/arXiv.2306.02552

Xiao, Y.; Cheng, Y.; Fu, J.; Wang, J.; Li, W.; and Liu, P. 2023. How Far Are LLMs from Believable AI? A Benchmark for Evaluating the Believability of Human Behavior Simulation. arXiv. doi.org/10.48550/arXiv.2312.17115

Xu, N.; Zhang, Q.; Du, C.; Luo, Q.; Qiu, X.; Huang, X.; and Zhang, M. 2025. Human-Like Conceptual Representations Emerge from Language Prediction. arXiv. doi.org/10.48550/arXiv.2501.12547

Xu, X.; Yao, B.; Dong, Y.; Gabriel, S.; Yu, H.; Hendler, J.; Ghassemi, M.; Dey, A. K.; and Wang, D. 2024. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8(1): 31. doi.org/10.1145/3643540

Yetman, C. C. 2025. Representation in Large Language Models. arXiv. doi.org/10.48550/arXiv.2501.00885

Zednik, C., and Boelsen, H. 2022. Scientific Exploration and Explainable Artificial Intelligence. *Minds and Machines* 32: 219-239. doi.org/10.1007/s11023-021-09583-6

Zhang, J.; Wang, J.; Li, H.; Shou, L.; Chen, K.; You, Y.; Xie, G.; Gong, X.; and Zhou, K. 2025. Train Small, Infer Large: Memory-Efficient LoRA Training for Large Language Models. arXiv. doi.org/10.48550/arXiv.2502.13533