Journal of Experimental Psychology: Animal Learning and Cognition

“Captured” by Centaur: Opaque Predictions or Process Insights?

Phillip H. Kieval and Cameron Buckner

Online First Publication, September 29, 2025. https://dx.doi.org/10.1037/xan0000410

CITATION

Kieval, P. H., & Buckner, C. (2025). “Captured” by centaur: Opaque predictions or process insights? *Journal of Experimental Psychology: Animal Learning and Cognition*. Advance online publication. https://dx.doi.org/10.1037/xan0000410

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology:
Animal Learning and Cognition

© 2025 American Psychological Association
ISSN: 2329-8456

https://doi.org/10.1037/xan0000410

OUTLOOK

“Captured” by Centaur: Opaque Predictions or Process Insights?

Phillip H. Kieval and Cameron Buckner
Department of Philosophy, University of Florida

Summary: Binz et al. (2025) describe several ways that *Centaur*—a new computational model that “captures” human behavior better than alternatives—can help develop a new unified theory of cognition. In this commentary, we evaluate several of these roles in light of recent achievements and empirical data, recommending increasingly explicit scrutiny of the various modeling roles that Centaur might play in developing new explanatory theories of human cognition.

Centaur is built by fine-tuning Llama—a pretrained large language model (LLM)—on a massive, curated data set called Psych-101, which contains millions of human behavioral choices in psychological experiments. Centaur out-predicts the base Llama model, several other variants, and domain-specific cognitive models on held-out participant responses. This performance generalizes to at least some new problem structures, altered cover stories, and domains outside the fine-tuning set.

Binz et al.’s talk of “capturing” human behavior invites misinterpretation, by blurring the line between merely phenomenological modeling that predict patterns of behavior, and process modeling that depicts internal mechanisms which explain those patterns by producing them in systematic ways. “Capturing” is ambiguous between these two aims, but critiques assuming that Centaur is intended as an explanatory process model might thereby miss the mark. We believe that Centaur is most charitably construed as a phenomenological model capable of aiding in process model selection by furnishing a high ceiling of predictability on previously observed patterns of human behavior.

Phenomenological models are evaluated according to their adequacy to patterns in the phenomena they predict. They depict behavioral dispositions of their target system in terms of input–output relations, treating the cognitive system as a black box. Phenomenological models of human behavior can function as tools for guiding hypothesis formation and experimental design by summarizing patterns that a process model needs to explain. An ideally complete phenomenological model would summarize all and only the patterns relevant to a phenomenon (including those not yet observed), but of course no models are ideal. For some purposes, a high degree of phenomenal adequacy is sufficient, and we have never had a phenomenological model that was this adequate for such a wide range of human behaviors.

Centaur’s main contribution as a phenomenological model is supported by the authors’ invocation scientific regret minimization (SRM), which uses machine learning models to aid researchers in selecting a process model by estimating the ceiling of predictability for a given data set. SRM leverages mechanistically unconstrained machine learning models to estimate the explainable variance present in large datasets. Modelers then penalize candidate process models only for failing to accommodate data predicted by the reference model. Binz et al. (2025) used DeepSeek-R1 as a tool for synthetic ideation to formulate a heuristic decision-making strategy that could explain human behavioral data drawn from a multi-attribute decision-making experiment. They then formalized this strategy into an algorithmic process model and used Centaur as a black-boxed reference model to refine their process model. This example demonstrates (at least one of) Centaur’s intended functions; Centaur shows promise as a tool for process modelers by extending the scope of SRM to domains where extensive experimental data is relatively scarce.

Even so, Centaur’s phenomenal adequacy remains imperfect if it is unstable under generalization, as shown by failures to capture human behavior in the 11–20 money request game (Gao et al., 2025) and experimental manipulations designed to test short-term memory (Bowers et al., 2025). Granted, memory capacity is the standard example of a “performance” factor that might not discount a model’s phenomenal adequacy to an underlying cognitive competence, but even this rebuttal would require theory-driven arbitration, to decide what does and does not count as part of the underlying competence. In general, such evidence of instability under slight task alteration points to the inherent limitations of purely behavioral evidence of generalization in an overparameterized deep learning paradigm. We know that Centaur’s capacity to generalize is both impressive and unstable, and we cannot escape this fundamental tension by using more data and training to buy arbitrarily greater degrees of fit to previously observed human behavioral patterns. If generalization bounds are unpredictable, we might want evidence that the representations and computations Centaur uses to predict human behavioral responses will track those actually used by humans. We may thus need adequate process models to know when to trust a deep learning model’s capacity to generalize.

More ambitious interpretations of Centaur’s intended purpose as a process model are not unfounded; Binz et al.’s decision to evaluate Centaur’s representational alignment with human neural activity implies that they also view Centaur as at least a partial or candidate process model. However, the simple measures of neural alignment that they report are limited in their ability to establish Centaur as a

Jonathon D. Crystal served as action editor.

Phillip H. Kieval https://orcid.org/0000-0001-5369-0322
Cameron Buckner https://orcid.org/0000-0003-0611-5354
Correspondence concerning this article should be addressed to Phillip H. Kieval, Department of Philosophy, University of Florida, 306 Griffin-Floyd Hall, Gainesville, FL 32611-8545, United States. Email: pkieval@ufl.edu

partial process model. They fit a regularized linear regression model to predict functional magnetic resonance imaging data from Centaur's internal representations, finding improved alignment compared to base Llama, a randomly-initialized architecture-matched transformer, and cognitive models. The fact that a linear model can predict neural activity from Centaur's layer-wise internal representations does not suffice to show that Centaur is a strong candidate process model. Comparison with a randomly-initialized control rules out that improved alignment is merely an input perturbation effect between the model's high effective dimensionality and the shared complex input with human participants. However, it does not confirm that the model uses this information induced by fine-tuning to generate its behavioral predictions. Ablation studies have revealed that transformers generally can contain significant amounts of redundant information induced by training or fine-tuning, the removal of which does not much affect the model's predictions.

Still, there are general reasons to expect that even mechanistically unconstrained machine learning models might independently develop and deploy some of the same representations that humans use to solve similar tasks. Recent evidence, for example, suggests that overparameterized transformer models are biased toward discovering simpler solutions to prediction problems. Despite their unrestricted flexibility, training LLMs well beyond the empirical loss ceiling can cause these models to converge on highly generalizable solutions with lower effective dimensionality (Wilson, 2025). From this perspective, deep learning can be viewed as a form of data compression, and the most efficient compression of data may be a causal model of the data generating process. Centaur's fine-tuning period was probably insufficiently long to induce this effect itself, but perhaps it was sufficient to recruit a sparser set of latent human-aligned representations from the base Llama's compression space to generate its predictions of human behavior. Yet, mere linear decodability is not sufficient to show that this is more than conjecture. Ruling out deflationary explanations of neural alignment requires mechanistic interventions to confirm which representations the model in fact uses to generate its predictions (for a review of work in this area, see Millière & Buckner, 2026).

Interventionist methods use hypothesis-driven alterations of specific elements in a model's activation space to verify that those elements play a causal role like the one they are theorized to play in human cognition. Of course, hopes here should be tempered; mechanistic interpretability methods are new and may have unknown limitations, and early results have sometimes revealed that deep learning models acquire heaps of unsystematic heuristics or solutions which are systematic, but which differ in their generalization profile from human cognition. For present purposes, our point is that if Centaur is to serve as something stronger than a phenomenological model for previously gathered human behavioral data, modelers should supplement purely behavioral methods with mechanistic interpretability experiments. Doing so might provide an additional source of evidence that could be leveraged to help decide when Centaur's ability to generalize to new situations can be relied upon as a proxy for novel human data.

To summarize, there are at least three ways to construe Centaur as a step toward a unified theory of human cognition: (a) most modestly, as an estimate of predictable variance in previously collected human data included in the fine-tuning set, (b) somewhat less modestly, as an estimate of statistics like predictable variance, effect sizes, and power for novel experiments not included in the fine-tuning set, and (c) most ambitiously, as itself a partial or candidate process model. The first use is the best supported (but least interesting), and the latter two are more precarious and intertwined than we might have initially supposed. Moving forward, Centaur's creators, consumers, and critics should be explicit about their construal of its interpretation and intended purpose.

References

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., Griffiths, T. L., Haridi, S., Jagadish, A. K., Ji-An, L., Kipnis, A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M., … Schulz, E. (2025). A foundation model to predict and capture human cognition. *Nature*, *644*(8078), 1002–1009. https://doi.org/10.1038/s41586-025-09215-4

Bowers, J. S., Puebla, G., Thorat, S., Tsetsos, K., & Ludwig, C. J. H. (2025, July). *Centaur: A model without a theory*. https://doi.org/10.31234/osf.io/v9w37_v3

Gao, Y., Lee, D., Burtch, G., & Fazelpour, S. (2025). Take caution in using LLMs as human surrogates. *Proceedings of the National Academy of Sciences of the United States of America*, *122*(24), Article e2501660122. https://doi.org/10.1073/pnas.2501660122

Millière, R., & Buckner, C. (2026). Interventionist methods for interpreting deep neural networks. In G. Piccinini (Ed.), *Neurocognitive foundations of mind* (pp. 190–221). Oxford University Press.

Wilson, A. G. (2025). *Deep learning is not so mysterious or different*. https://arxiv.org/abs/2503.02113

Received August 21, 2025
Revision received August 27, 2025
Accepted August 28, 2025 ■