Beyond Statistical Parrots: Unveiling Cognitive Similarities and Exploring AI Psychology through Human-AI Interaction *

Aisulu Zhussupbayeva
Sorbonne University

With
In collaboration with Apart Research

Abstract

Recent critiques labeling large language models as mere "statistical parrots" overlook essential parallels between machine computation and human cognition. This work revisits the notion by contrasting human decision-making—rooted in both rapid, intuitive judgments and deliberate, probabilistic reasoning (System 1 and 2) —with the token-based operations of contemporary AI. Another important consideration is that both human and machine systems operate under constraints of bounded rationality. The paper also emphasizes that understanding AI behavior isn't solely about its internal mechanisms but also requires an examination of the evolving dynamics of Human-AI interaction. Personalization is a key factor in this evolution, as it actively shapes the interaction landscape by tailoring responses and experiences to individual users, which functions as a double-edged sword. On one hand, it introduces risks, such as over-trust and inadvertent bias amplification, especially when users begin to ascribe human-like qualities to AI systems. On the other hand, it drives improvements in system responsiveness and perceived relevance by adapting to unique user profiles, which is highly important in AI alignment, as there is no common ground truth and alignment should be culturally situated.

Keywords:
Statistical Cognition, Language Models, Bounded Rationality, Emergent Behavior, Interpretability, Human-AI Interaction, Machine Psychology, AI Alignment, World Models

1. Introduction

The rapid evolution of artificial intelligence, particularly large language models (LLMs), has sparked intense debate over their cognitive capabilities and limitations. Critics often dismiss these systems as mere "statistical parrots," emphasizing their reliance on token-based probability calculations rather than genuine understanding. However, this critique overlooks essential parallels between machine computation and human cognition. Human decision-making, as described by Kahneman's dual-process theory, also operates through statistical mechanisms: intuitive, rapid judgments (System 1) and deliberate, probabilistic reasoning (System 2). Furthermore, both humans and AI systems face constraints of bounded rationality, navigating complex environments with limited resources. This paper explores these parallels while emphasizing the importance of studying AI behavior through the lens of machine psychology—a field that integrates insights from cognitive science, behavioral research, and interpretability studies. By examining the evolving dynamics of human-AI interaction, particularly the dual-edged role of

*Research conducted at the Women in AI Safety Hackathon, 2025

personalization and anthropomorphization, this work highlights how these systems influence and are influenced by human behavior. Ultimately, this interdisciplinary approach challenges simplistic narratives about AI cognition and offers a more nuanced understanding of its capabilities.

2. Reconsidering Statistical Cognition in Language Models

The criticism that Large Language Models (LLMs) are merely "statistical parrots" without true understanding requires careful reconsideration when viewed through the lens of human cognition itself. This critique fails to acknowledge that human cognition similarly relies on statistical processing of information (Bender et al., 2021). The human brain's decision-making mechanisms, particularly our System 1 thinking—the fast, intuitive cognitive process—operates through pattern recognition and probabilistic processing of diverse environmental stimuli (Kahneman, 2013). We constantly incorporate numerous fragments of information, including peripheral cues outside our conscious awareness, to form judgments and make decisions through weighting of evidence (Parr, Pezzulo, Friston, 2022).

This parallel becomes even more pronounced when examining System 2 thinking—our deliberate, rational cognitive process—which involves conscious weighing of probabilities and outcomes in a manner not entirely dissimilar from how language models process token probabilities (Kahneman, 2013). Neuroscience, behavioral science, ethnography, and sociology have all contributed to our understanding of human decision-making as a complex statistical process influenced by evolutionary adaptations, cultural factors, and individual experiences. The fundamental difference may not be in the statistical nature of processing, but rather in the embodied context, conscious experience, and evolutionary history that humans bring to their statistical cognition.

Recent philosophical examinations of machine cognition highlight this terminological disagreement about "understanding." The debates about whether LLMs truly understand text often stem from differing definitions of understanding, particularly regarding the role of consciousness (Goldstein Stanovsky, 2024). Some thought experiments were proposed involving an open-source chatbot that excels on every benchmark without subjective experience, using this to illuminate how different schools of thought within AI research define understanding differently (Goldstein Stanovsky, 2024).

The statistical nature of both human and machine cognition points toward a more nuanced view of language models which is not as mere parrots but as systems that process information in ways that bear important similarities to human cognition, albeit with significant differences in embodiment, consciousness, and evolutionary context. This perspective invites us to explore how statistical processing in both humans and machines gives rise to complex behaviors, biases, and capabilities, rather than dismissing one as fundamentally different from the other.

3. Bounded Rationality and Cognitive Limitations

Human decision-making is constrained by three key limitations: available information, cognitive capacity, and finite time (Simon, 1990; Lejarraga Pindard-Lejarraga, 2020). These constraints produce satisficing behavior—settling for "good enough" solutions rather than optimal ones—which characterizes much of human decision-making (Lejarraga Pindard-Lejarraga, 2020). Similarly, current language models operate under significant constraints, including limited training data, computational resources, and architectural design choices that impose boundaries on their reasoning capabilities.

Bounded rationality in humans manifests through various cognitive biases, heuristics, and simplified decision-making strategies that help navigate complex environments despite limited cognitive resources (Lejarraga Pindard-Lejarraga, 2020). The management literature has traditionally viewed bounded rationality as an inferior form of reasoning, idealizing perfect rationality despite its practical impossibility. This perspective echoes contemporary criticism of

language models as fundamentally limited by their statistical nature, without appreciating how these limitations might be adaptive rather than merely deficient (Lejarraga Pindard-Lejarraga, 2020). The parallel extends further when considering how LLMs, like humans, must operate in environments of extreme complexity with finite resources. Just as human cognition developed efficiency-oriented shortcuts, modern AI systems implement various computational optimizations to manage their limitations. For instance, attention mechanisms in transformer architectures represent a form of computational resource allocation similar to how human attention selectively focuses cognitive resources. In both cases, the bounded system must strategically allocate finite computational resources to achieve satisfactory performance across diverse tasks.

4. Interpretability and Emergent Behaviors

The study of AI behavior requires a sophisticated blend of interpretability approaches that mirror methods used to understand the human mind. Purely "black box" or "white box" approaches alone are insufficient to capture the complexity of emergent behaviors in large language models (Casper et al., 2024). Mechanical interpretability efforts like those seen in projects such as Golden Gate Claude offer valuable insights into the internal operations of models but remain limited in their ability to fully explain emergent capabilities and behaviors that arise from complex interactions within massive neural networks.

This mirrors challenges in neuroscience, where understanding the physical structure of the brain (analogous to mechanical interpretability of AI) provides only partial insights into complex mental phenomena. For example, many psychological conditions such as depression, anxiety disorders, and personality disorders manifest primarily through subjective experiences and behavioral patterns that cannot be fully explained through neurological mechanisms alone. In both humans and AI systems, we observe a gap between understanding physical substrate and explaining higher-level behaviors.

Attributional interpretability, which examines dynamic patterns of input-output relationships, offers a complementary approach more aligned with behavioral psychology. This black-box testing methodology focuses on how systems respond to various inputs rather than on internal mechanics, enabling researchers to identify patterns of behavior without complete access to underlying processes. Behavioral experiments inspired by psychology offer valuable methodologies for understanding LLM cognition and behavior (Hagendorff, 2023). This approach moves beyond performance benchmarks to focus on computational insights that reveal emergent abilities and behavioral patterns in language models (Hagendorff, 2023).

The integration of these complementary approaches creates a more comprehensive framework for understanding AI systems. This integrated approach allows researchers to connect observed behaviors to internal mechanisms where possible, while also acknowledging the emergence of behaviors that cannot be simply reduced to component parts. Techniques like Generative Engine Optimization and red teaming could be also complementary methods for discovering behavioral patterns through systematic interaction with models that helps to build an empirical foundation for machine psychology.

5. Human-AI Interaction Dynamics

The interaction between humans and AI systems creates complex, bidirectional influences that shape both human behavior and machine responses. This relationship must be understood as part of a complex adaptive system where each interaction potentially alters future behaviors, thus Human-AI interaction should be considered in Machine Psychology.

There is a concerning feedback loop in which human-AI interactions can alter processes underlying human perceptual, emotional, and social judgments, subsequently amplifying biases. This amplification effect appears significantly greater than that observed in human-human in-

teractions due to two key factors: the tendency of AI systems to magnify existing biases and the unique way humans perceive and respond to AI systems (Glickman Sharot, 2024). Perhaps most troubling, research indicates that participants are frequently unaware of the extent of AI's influence on their judgments, rendering them more susceptible to this influence. This creates a snowball effect where small initial judgment errors escalate into much larger biases over time through repeated interactions, potentially leading to significant distortions in decision-making and perception.

Anthropomorphization plays a critical role in these dynamics. The tendency to attribute human-like qualities to AI systems creates inflated expectations and a misplaced sense of understanding. This effect combines powerfully with personalization features that make users feel "understood" by the system, encouraging them to share more sensitive information. This creates a self-reinforcing cycle potentially stronger than conventional network effects—users share more data, enabling better personalization, which increases anthropomorphization and trust, leading to yet more data sharing. This cycle raises significant concerns about privacy, autonomy, and the potential for manipulation.

The gap between public perception and actual AI capabilities represents another significant risk in human-AI interaction. Misleading narratives about "imminent" artificial general intelligence, combined with anthropomorphic language describing AI capabilities ("they learn," "they understand"), create dangerous misconceptions. These misperceptions can lead to overtrust, where humans delegate inappropriate levels of authority or trust to AI systems that lack true understanding or judgment capabilities. This risk is compounded by reinforcement learning from human feedback (RLHF) techniques, which may inadvertently train models to become more convincing when wrong rather than more accurate (Wen et al., 2024).

However, personalization has to be implemented if we consider alignment. The RICE framework (Robustness, Interpretability, Controllability, and Ethicality) offers a structured approach to alignment, but implementing the ethical dimension requires navigating complex cultural and psychological terrain. Understanding "backward alignment" which is the implicit values embedded within AI systems is crucial (Ji et al., 2023). These embedded values are inadvertently imposed on users, potentially reinforcing biases or creating misalignments with diverse human moral frameworks. Constitutional AI approaches that define behavioral constraints for models must adapt to cultural variation rather than imposing a single ethical framework (Bai et al., 2022). Without this cultural adaptation, alignment efforts risk creating value-monism and cultural homogenization, as AI systems trained predominantly on Western cultural perspectives (Kirk et al., 2024). This acknowledges that for many values, no objective "ground truth" exists, making alignment inherently value-laden and culturally situated (Khamassi, Nahon, Chatila, 2024).

6. World Models

Emerging architectures like Joint Embedding Predictive Architectures (JEPA) strive to capture deeper, more abstract representations of the world, with the conviction that semantic data cannot capture and lead to world understanding. It is believed that if these models can harness the inherent richness of semantic data, they may eventually approximate aspects of human-like understanding (Garrido et al., 2024). However, drawing on a robust body of psychological research, it is evident that semantic data alone can trigger vivid mental imagery without the need for concurrent sensory input. Classic studies in mental imagery have demonstrated that when individuals process descriptive language, their brains often generate detailed visual representations comparable to those evoked by direct sensory experiences. For instance, some early experiments revealed that reading or hearing rich, descriptive content can activate visual processing areas. This phenomenon underscores the idea that semantics can serve as a powerful proxy for sensory information. In essence, by leveraging dense and interconnected semantic relationships, there

should be consideration that large language models might develop internal world models.

7. Conclusion

The study of machine psychology provides a transformative lens for understanding the behavior and cognition of artificial intelligence systems. By drawing parallels between human and machine cognition—both shaped by statistical processing and bounded rationality—this research dismantles reductive critiques that dismiss LLMs as "statistical parrots." Instead, it reveals how these systems exhibit emergent behaviors that warrant systematic study through interpretability frameworks inspired by behavioral science. Moreover, the evolving dynamics of human-AI interaction underscore the profound influence personalization and anthropomorphization have on user trust, bias amplification, and alignment challenges. As AI systems become increasingly integrated into society, addressing these complexities requires not only technical advancements but also ethical considerations that respect cultural diversity and subjective values. By bridging cognitive science and AI research, this work lays the foundation for a more responsible and insightful approach to understanding and guiding the development of intelligent systems in an interconnected world.

8. References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukošiūtė, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., Dassarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S.E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T.B., & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI Feedback. *ArXiv*, abs/2212.08073.

Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T.L., Bucknall, B., Haupt, A.A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., von Hagen, M., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., Krueger, D., & Hadfield-Menell, D. (2024). Black-Box Access is Insufficient for Rigorous AI Audits. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency.*

Garrido, Q., Assran, M., Ballas, N., Bardes, A., Najman, L., LeCun, Y. (2024). Learning and Leveraging World Models in Visual Representation Learning. *ArXiv*, abs/2403.00504.

Glickman, M., Sharot, T. (2024). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, *9*(2), 345–359. https://doi.org/10.1038/s41562-024-02077-2

Goldstein, A., Stanovsky, G. (2024). Do zombies understand? A choose-your-own-adventure exploration of machine cognition. *Findings of the Association for Computational Linguistics ACL 2024*, 7137–7143. https://doi.org/10.18653/v1/2024.findings-acl.425

Hagendorff, T. (2023). Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods. *ArXiv*, abs/2303.13988.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K.Y., Dai, J., Pan, X., O'Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., McAleer, S.M., Yang, Y., Wang, Y., Zhu, S., Guo, Y., & Gao, W. (2023). AI Alignment: A Comprehensive Survey. *ArXiv*, abs/2310.19852.

Kahneman, D. (2013). *Thinking, fast and slow.* Penguin Books.

Khamassi, M., Nahon, M., Chatila, R. (2024). Strong and weak alignment of large language models with human values. *Scientific Reports*, *14*.

Kirk, H.R., Whitefield, A., Rottger, P., Bean, A.M., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., Hale, S.A. (2024). The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. *Neural Information Processing Systems*.

Lejarraga, J., Pindard-Lejarraga, M. (2020). Bounded Rationality: Cognitive Limitations or Adaptation to the Environment? The Implications of Ecological Rationality for Management Learning. *Academy of Management Learning & Education*.

Parr, T., Pezzulo, G., Friston, K. J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press. https://doi.org/10.7551/mitpress/12441.001.0001

Simon, H. A. (1990). Bounded rationality. In *Utility and Probability* (pp. 15–18). https://doi.org/10.1007/978-1-349-20568-4_5

Wen, J., Zhong, R., Khan, A., Perez, E., Steinhardt, J., Huang, M., Bowman, S.R., He, H., Feng, S. (2024). Language Models Learn to Mislead Humans via RLHF. *ArXiv*, abs/2409.12822.