The International Journal of Indian Psychology
ISSN 2348-5396 (Online) | ISSN: 2349-3429 (Print)
Volume 12, Issue 2, April- June, 2024
DIP: 18.01.288.20241202, DOI: 10.25215/1202.288
https://www.ijip.in

Research Paper

Understanding "Minds" of the Machines: Exploring the Role of Psychology in Fostering the Development of Artificial Intelligence in Contemporary Times

Priyanshi Jangra[1]*

ABSTRACT

Artificial Intelligence (AI) has become an integral aspect of human lives in contemporary times. It has applications in virtually every field, like healthcare, education, and security. While it continues to flourish, AI is not flawless. The increasing complexity of AI systems has led to the emergence of a critical issue known as the "black box problem" wherein the inner workings of these systems remain opaque and difficult to interpret. This opacity hampers our understanding of AI decision-making processes and raises concerns regarding accountability, transparency, and trustworthiness. Furthermore, in certain areas, artificial intelligence is still far behind human cognition. Psychology, which has its roots in human cognition, behaviour, and emotion, offers a sophisticated lens through which AI devices might be given more transparency and clarity. This paper explores the interdisciplinary field of artificial cognition, which assimilates cognitive psychological principles in modern AI. Through a comprehensive analysis of artificial cognition techniques and their implications for artificial intelligence, this paper aims to contribute to the ongoing discourse on mitigating the black box problem and advancing the responsible development and deployment of AI technology. The larger goal served by this paper is the presentation of how the discipline of psychology is relevant and even necessary in a world dominated by AI.

Keywords: *Artificial Intelligence, Explainable Artificial Intelligence, Artificial Cognition, Black-Box Problem, Cognitive Psychology*

Artificial intelligence (AI) can be defined as the capability of a machine to simulate the human brain and exhibit the capacity to think, learn, and act like humans. It comprises the techniques that enable the machine to exhibit human-like intelligence with minimal intervention from a human being. The emergence of Artificial Intelligence (AI) can be linked to the works of the three giants- Charles Babbage (1791-1871), Augusta Ada King (Ada Lovelace) (1815–1852), and Alan Mathison Turing (1912 - 1954)- who dominated the early development of this field (Grzybowski, 2024). Some of the landmark discoveries in this field include the Turing test, the General Problem Solver program developed by Herbert Simon, Cliff Shaw, and Allen Newell, the ELIZA computer program, created between 1964 and 1966 by Joseph Weizenbaum, and so on (Haenlein & Kaplan, 2019).

[1]Department of Psychology, Daulat Ram College, University of Delhi, Delhi, India
*Corresponding Author

Received: March 24, 2024; Revision Received: June 06, 2024; Accepted: June 10, 2024

In the present time, AI is completely intertwined with the lives of individuals and society at large. AI applications are used to improve business efficiency (e.g., recommendation systems, chatbots), assist in daily life (e.g., voice command-based digital assistants), and automate complex operations (e.g., autopilot in vehicles) (Kellogg et al., 2020). In healthcare, AI-driven diagnostic tools leverage insights from cognitive psychology to enhance clinical decision-making, while virtual reality-based therapies draw upon principles of behavioral psychology to facilitate interventions for mental health disorders. Similarly, in educational settings, AI tutors harness principles of psychology to tailor learning experiences to individual student needs, fostering personalised and adaptive learning pathways.

The advancements in AI largely stemmed from making a system mimic a human brain. While that has been working out for the scientific community so far, it is not without any drawbacks. This paper aims to highlight some of the problems encountered by AI developers and computer scientists and highlight how the discipline of psychology has been of help in tackling those problems. The larger goal of this work is to suggest how psychologists will stay relevant even in an "algorithmically infused society" (Wagner et al., 2021).

The first section of the paper describes the black-box problem in AI, i.e., the opacity of the decision-making process of the AI systems, and subsequently suggests how the discipline of psychology has helped in tackling it. The second section of the paper describes how AI is still not superior to humans in some areas, and how psychology offers solutions for the same.

The Black-Box Problem

Artificial intelligence (AI) is an all-encompassing term for the field focused on studying and building intelligent systems. A subset of AI is machine learning (ML) which is capable of 'learning from mistakes' such that the computer program's performance of a particular type of task improves by learning from previous computations and extracting regularities from the data (Janiesch et al., 2020). Patterns in the data can be automatically found by machine learning.

Similarly, deep learning is a kind of machine learning that simulates how the human brain learns by using artificial neural networks. It uses deep neural networks (DNN) which are organised as deeply convoluted networks. Deep learning simulates the deep layered networks of the human brain. Since DNNs mimic the human brain, they are essentially opaque, just like our brains. For instance, with human beings, reasoning about decisions is a highly flexible and ad hoc process. People are often unable to appropriately report on their implicit attitudes or cognitions since they are typically unaware of them. Because of this, it can be challenging for many people to critically examine their conscious mental states and then verbally describe and characterise the conclusions of that examination. To put it another way, we can learn nothing about the process by which our conscious thoughts originated (Korteling et al., 2021). A similar problem is being encountered by computer scientists with AI, which is called the black-box problem. Often, an AI model's user or developer cannot explain the model's decisions since these models refine themselves autonomously and with nuances beyond human comprehension and computation (Taylor & Taylor, 2021). Thus, the problem-solving approach of AI remains incomprehensible.

Compared to previous technologies, transparency is a bigger concern for AI since it operates in a complicated, multi-layered manner, with challenging logic to comprehend (Glikson & Woolley, 2020). The black box problem in AI is quite significant considering the vast applications of AI in the healthcare system, education system, and so on.

The need for explainability and transparency is also connected to the idea of trust: we want to be sure that a proposed technical solution "does the right thing," without inflicting harm (Liem et al., 2018). Building trust in new technologies requires transparency, which measures how much people comprehend a system's inner logic or workings.

There is a need to increase the transparency of AI not just on a societal level, but also on an individual level. This can be highlighted by the study done by Liu (2021) which assessed the extent to which transparency of AI reduces the sense of uncertainty among users and subsequently, increases their trust. In the study, 491 participants were made to interact with an online website whose AI system was manipulated on agency locus (human-made rules vs. machine-learned rules), transparency (no vs. placebic vs. real explanations), and task (detecting fake news vs. assessing personality).

The findings revealed that participants felt more informed about the AI system's decision-making process when they were aware that it was designed to obey human-made rules, and thus, had a lower sense of uncertainty and higher trust. Furthermore, even placebic transparency on the part of the system decreased the sense of uncertainty on the part of the users. These findings highlight the psychological need for transparency in human-machine interaction.

To tackle this issue, a class of models called Explainable Artificial Intelligence (XAI) have attempted to open the 'black box' but have proved limited in their success (von Eschenbach, 2021). Furthermore, additional problems are encountered when an AI model's decision-making processes are made to be inferred by another AI model. To this problem, Taylor and Taylor (2021) have suggested the plausibility of drawing upon psychological experimental methods and principles to increase the explainability of AI. They have referred to this area of research as Artificial Cognition (AC), i.e., the use of experimental psychology for understanding, evaluating, and explaining machine learning algorithms.

The rationale behind this proposal stems from the common issues faced by AI researchers and cognitive psychologists. Years ago when cognitive psychology was still budding, the problem faced by psychologists was the lack of understanding of how mental processing is carried out in human beings. The primary focus of cognitive psychology is the advanced mental processes of human cognition, such as the degree of reasoning, emotion, motivation, and decision-making. The assumption that humans are information-processing systems forms the basis of most contemporary conceptualizations of human thought. Perceive, attend, memorise, and think. All of these activities tamper with people's mental information. Through the use of contemporary techniques in cognitive psychology, researchers can monitor how individuals absorb information and, as a result, gain insight into the types of information processing that enable intelligent behaviour (Saariluoma & Karvonen, 2021). Likewise, the same approach can be applied to the "minds" of the machines.

This analogy gave impetus to solving the problem of understanding the black-box processing in AI using experimental methods from psychology and trying the understand the

algorithms the same way psychologists analyse mental processes. This kind of work is done in the Psychlab which is a virtual psychological laboratory designed by DeepMind.

Psychlab creates a similar framework to test the cognitive capacities of an AI agent alongside human participants in the virtual environment of DeepMind Lab, much like how human psychology tests might be set up in a clinical setting. This enables direct comparisons between deep reinforcement learning agents and humans on tasks taken straight from visual psychophysics and cognitive psychology (Leibo et al., 2018). Examples of such tasks are visual search, random dot motion discrimination, multiple object tracking, etc. The Psychlab for machine learning agents offers an opportunity to gain insights into the decision-making process of the algorithms.

Similarly, Ritter et al. (2017) used the knowledge about shape bias, i.e., the tendency to categorise objects according to shape rather than colour, size, or texture, in children from developmental psychological research to study DNN behaviour. They found a similar shape bias in DNN as it is seen in human beings.

Diehl and de Vries (2020) studied holistic and featural face processing in neural networks using psychological testing methods.

Considering how recent this field is, very few researchers have attempted this kind of work. Nonetheless, to advance AI, more and more researchers must begin to study this area.

Artificial Cognition

Artificial cognition is a relatively new area of study, and thus, there is a lack of consensus among researchers regarding its definition and goals. While Taylor and Taylor (2021) aim to further XAI via the work done in artificial cognition, Siemens et al. (2022) have defined AC as the field that studies those aspects of AI that are cognition-like and share overlapping capabilities with humans. Their goal is to comprehend the nature of cognitive processes that artificial intelligence (AI) systems carry out that are similar to, superior to, or even alter human cognition. Given that AI is still behind human cognition and that it has the potential to grow, this kind of work is essential. For instance, considering that AI is programmed by humans and can only adhere to and carry out the instructions that are programmed into it, the fundamental programming cannot be changed by it on its own (Tariq et al., 2022). To do such, a human agent is required.

This approach to AC invites psychologists to contribute to the development of AI by making cognitive architectures that would eventually assist in improving the functioning of AI. One such cognitive architecture is Connectionist Learning with Adaptive Rule Induction On-line (CLARION).

CLARION is a computational cognitive architecture that provides a framework for understanding human cognition, particularly the interaction between explicit and implicit learning processes. By modeling the human mind's dual-process nature, CLARION has offered insights into how humans learn from both conscious reasoning and unconscious, intuitive processes. This architecture has had implications for AI by inspiring the development of more human-like learning algorithms. For instance, CLARION-inspired models have been applied in AI systems to improve decision-making, autonomous agents, and adaptive learning algorithms (Sun, 2006). By incorporating principles from CLARION,

AI researchers aim to create more flexible, adaptive systems capable of learning from experience and making decisions in complex, dynamic environments.

Another cognitive architecture that has been extensively employed within AI is Adaptive Control of Thought-Rational (ACT-R).

ACT-R also offers a computational framework for understanding human cognition and behavior. It breaks down the mind into components such as declarative and procedural memory, working memory, and a production system, simulating how humans process information and make decisions. In the realm of AI, ACT-R serves as a valuable tool for developing cognitive models and intelligent systems that emulate human-like reasoning and learning. For example, ACT-R-inspired models have been used in AI applications such as cognitive tutoring systems, human-robot interaction, and adaptive training systems (Anderson et al., 2004).

The concept of artificial cognition is similar to that of Psychological artificial intelligence which refers to the use of psychological theories and principles in the designing and development of AI algorithms, particularly to make these algorithms 'smart' and 'transparent' (Gigerenzer, 2023). Psychological AI makes use of psychological theories of mind while making the algorithms.

This approach was also utilised by the participants in the Choice Prediction Competition in 2015 wherein the superiority of psychological AI over conventional machine learning models was demonstrated. The researchers were given a task to propose a model that could forecast human decisions in situations involving uncertainty and risk.

The results of the competition showed that the models created by economists and psychologists using a collection of heuristics that had been identified in the psychology literature tended to outperform the ML techniques (Ho & Griffiths, 2022; Noti et al., 2016). Thus, ML algorithms could be enhanced by incorporating the models of behavioural sciences.

However, there isn't much research done in this field either. Although some researchers have conducted comparable studies, we find that psychologists remain detached from it.

CONCLUSION

This paper identified new research directions where psychologists can contribute to the development of AI- artificial cognition and psychological artificial intelligence. However, despite the need for psychologists in these fields has been realised, their contribution remains less.

It is worth noting that integrating psychologists into the process of solving the black box problem in AI is not only beneficial but also essential for the development of responsible and ethical artificial intelligence systems. Moreover, psychologists can contribute to the design of user-friendly interfaces and interaction modalities that enhance user comprehension and trust in AI technologies. Psychologists bring unique insights into human cognition, behavior, and decision-making processes, which are fundamental to understanding how AI systems interact with users and society at large. By collaborating with psychologists, computer scientists can gain a deeper understanding of the underlying

mechanisms driving AI decision-making, thereby facilitating the development of more transparent, interpretable, and trustworthy AI systems.

The applications of psychology in AI have vast implications for the discipline of psychology as well considering the avenues of research it opens up. Now, apart from the traditional routes to the application of psychological theories and principles in various settings, like clinical, educational, organisational, etc., it can be practiced on digital systems. However, from a critical point of view, this approach also changes the unit of analysis for the discipline. What sets psychology apart from other social scientific disciplines has been its focus on the individuals, as opposed to groups, cultures, or societies. Analysing the "minds" of machines would challenge the unit of analysis of psychology.

Nonetheless, such work has reciprocal effects. The computational models developed by applying psychological experimental methods do, in the end, provide us with a deeper understanding of the human psyche.

REFERENCES

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4), 1036–1060.

Diehl, A., & de Vries, I. A. (2020). Cognitive Psychology for Black Box Models. Gigerenzer, G. (2023). Psychological AI: Designing algorithms informed by human psychology. *Perspectives on Psychological Science*.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660.

Grzybowski, A., Pawlikowska–Lagod, K., & Lambert, W. C. (2024). A history of artificial intelligence. *Clinics in Dermatology*.

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, *61*(4), 5-14.

Ho, M. K., & Griffiths, T. L. (2022). Cognitive science as a source of forward and inverse models of human decisions for robotics and control. *Annual Review of Control, Robotics, and Autonomous Systems*, *5*, 33-53.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685-695.

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, 14, 366-410.

Korteling, J. H., van de Boer-Visschedijk, G. C., Blankendaal, R. A., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human-versus artificial intelligence. *Frontiers in artificial intelligence*, *4*, 622364.

Leibo, J. Z., d'Autume, C. D. M., Zoran, D., Amos, D., Beattie, C., Anderson, K., ... & Botvinick, M. M. (2018). Psychlab: a psychology laboratory for deep reinforcement learning agents. *arXiv preprint arXiv:1801.08116*.

Liem, C. C., Langer, M., Demetriou, A., Hiemstra, A. M., Sukma Wicaksana, A., Born, M. P., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. *Explainable and interpretable models in computer vision and machine learning*, 197-253.

Liu, B. (2021). In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human–AI interaction. *Journal of Computer-Mediated Communication*, *26*(6), 384-402.

Noti, G., Levi, E., Kolumbus, Y., & Daniely, A. (2016). Behavior-based machine-learning: A hybrid approach for predicting human decision making. *arXiv preprint arXiv:2105.01160.*

Raisamo, R., Rakkolainen, I., Majaranta, P., Salminen, K., Rantala, J., & Farooq, A. (2019). Human augmentation: Past, present and future. *International Journal of Human-Computer Studies*, 131, 131-143.

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. In *International conference on machine learning*, 2940-2949.

Saariluoma, P., & Karvonen, A. (2021). The Psychology of Thinking in Creating AI. In *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)* (pp. 1-6). IEEE.

Siemens, G., Marmolejo-Ramos, F., Gabriel, F., Medeiros, K., Marrone, R., Joksimovic, S., & de Laat, M. (2022). Human and artificial cognition. *Computers and Education: Artificial Intelligence*, *3*, 100107.

Sun, R. (2006). The CLARION cognitive architecture: Extending cognitive modeling to social simulation. *Cognition and multi-agent interaction*, 79-99.

Sun, R. & Helie, S. (2013). Psychologically realistic cognitive agents: taking human cognition seriously. Journal of Experimental & Theoretical Artificial Intelligence, 25, 65-92.

Tariq, S., Iftikhar, A., Chaudhary, P., & Khurshid, K. (2023). Is the 'Technological Singularity Scenario' Possible: Can AI Parallel and Surpass All Human Mental Capabilities? *World Futures*, *79*(2), 200-266.

Taylor, J. E. T., & Taylor, G. W. (2021). Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin & Review*, *28*(2), 454-475.

Von Eschenbach, W.J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology*, 34, 1607-1622.

Wagner, C., Strohmaier, M., Olteanu, A., Kıcıman, E., Contractor, N., & Eliassi-Rad, T. (2021). Measuring algorithmically infused societies. *Nature*, *595*(7866), 197-204.

Zhao, J., Wu, M., Zhou, L., Wang, X., & Jia, J. (2022). Cognitive psychology-based artificial intelligence review. *Frontiers in Neuroscience*, *16*, 1024316.

Acknowledgment

The author(s) appreciates all those who participated in the study and helped to facilitate the research process.

Conflict of Interest

The author(s) declared no conflict of interest.

How to cite this article: Jangra, P. (2024). Understanding "Minds" of the Machines: Exploring the Role of Psychology in Fostering the Development of Artificial Intelligence in Contemporary Times. *International Journal of Indian Psychology*, *12*(2), 3261-3267. DIP:18.01.288.20241202, DOI:10.25215/1202.288