The International Journal of Indian Psychology
ISSN 2348-5396 (Online) | ISSN: 2349-3429 (Print)
Volume 11, Issue 3, July- September, 2023
DIP: 18.01.092.20231103, DOI: 10.25215/1103.092
https://www.ijip.in

Research Paper

Prediction Machines - Comparative Analysis of Theory of Mind Abilities in Machine Learning Models

Shannon J. Fernandes[1]*, Dr. Tejeshwar Dhananjaya[2]

ABSTRACT

With the advent of Machine Learning models in the current global markets, questions pertaining to artificial mental states and Artificial General Intelligence (AGI) come to the forefront. The current study seeks to find support for possible mental states by subjecting three different Machine Learning models - GPT 3.5 (Generative Pre-Trained Transformer), Bard, and GPT 3 - Ada to psychological tests concerning Theory of Mind (ToM). To discern another's mental state one must be capable of distinguishing oneself as an entity separate from the other. The three models were presented with the Strange Story Task and the Theory of Mind Scale. The idea of artificial ToM by observation and prediction has been posited. As hypothesized, GPT 3.5 significantly outperformed GPT 3 (Ada) and failed to outperform Bard. Further implications and limitations have been discussed.

Keywords: *Machine Learning, AI, Theory of Mind, AGI, Natural Language Processing, Cognition*

Theory of Mind (ToM) has served as a collaborative field of research, including areas such as psychology (and by extension cognitive psychology), philosophy, and neuroscience, to name a few. Theory of Mind refers to an agent's ability to represent another agent's mental state (Rabinowitz, et al., 2018). Apart from human's ability to perceive other mental states, theory of mind has been noted in animals as well, one noted example arrives from ravens who protect their collection when they realize they are being watched, as opposed to when they are not (Bugnyar, Reber, & Buckner, 2016). The sense of perceiving oneself or an object, through the lens of another, is at the core of ToM.

Over the past few decades, Machine Learning (ML) is beginning to contribute to the field of ToM as well (Rabinowitz, et al., 2018). A machine capable of displaying theory of mind abilities or consciousness as a whole, has long been hypothesized, and popularized by the Turing Machine (Bringsjord, Bello, & Ferrucci, 2003). It might appear that we are at the precipice of an era where theory of mind could be displayed in our existing ML models, a hypothesis this paper seeks to test, and with time, a measurable mental state in machines is likely to be observed. To the questions pertaining to the need for a psychological inquiry, Kosinski (2023), in this regard, implores the exploration of psychological sciences to better

[1]Indian Institute of Psychology & Research, Bangalore
[2]Indian Institute of Psychology & Research, Bangalore
*Corresponding Author

Received: May 24, 2023; Revision Received: July 20, 2023; Accepted: July 23, 2023

understand the behavior and mental states exhibited by models like GPT 3 (Generative Pre-Trained Transformer).

EMERGENCE & THEORY OF MIND

Emergence & Prediction Machines

The dissatisfaction with the dualist explanation of consciousness led to emergent theories of consciousness that focus primarily on the brain ranging from large networks to subcellular activities, where consciousness is perceived as an emergent property of the brain (Guevara, Mateos, & Velázquez, 2020). Emergence refers to complex or advanced processes or functions originating from seemingly simple processes or aggregate functions of the items or parts that form such a system (Feinberg, & Mallatt, 2020). It is fairly well observed in nature, be it in living organisms like ants or non-living beings (Johnson, 2001). Thus, the assumption posited is of neurons that can be held to the same standard as being fairly simple units that hold a binary outcome, but in a larger connected network, like that seen in the human brain, result in an emergent property we call consciousness.

Our phenomenological experience, especially that of being perceived as an agent by other agents (ToM), could be seen as a process that arises out of the several different computations that occur in the human brain. Given the fact that we observe theory of mind in different animals, as mentioned above, we could make a theoretical assumption of ToM serving as a predecessor to consciousness (Sebastián, 2016).

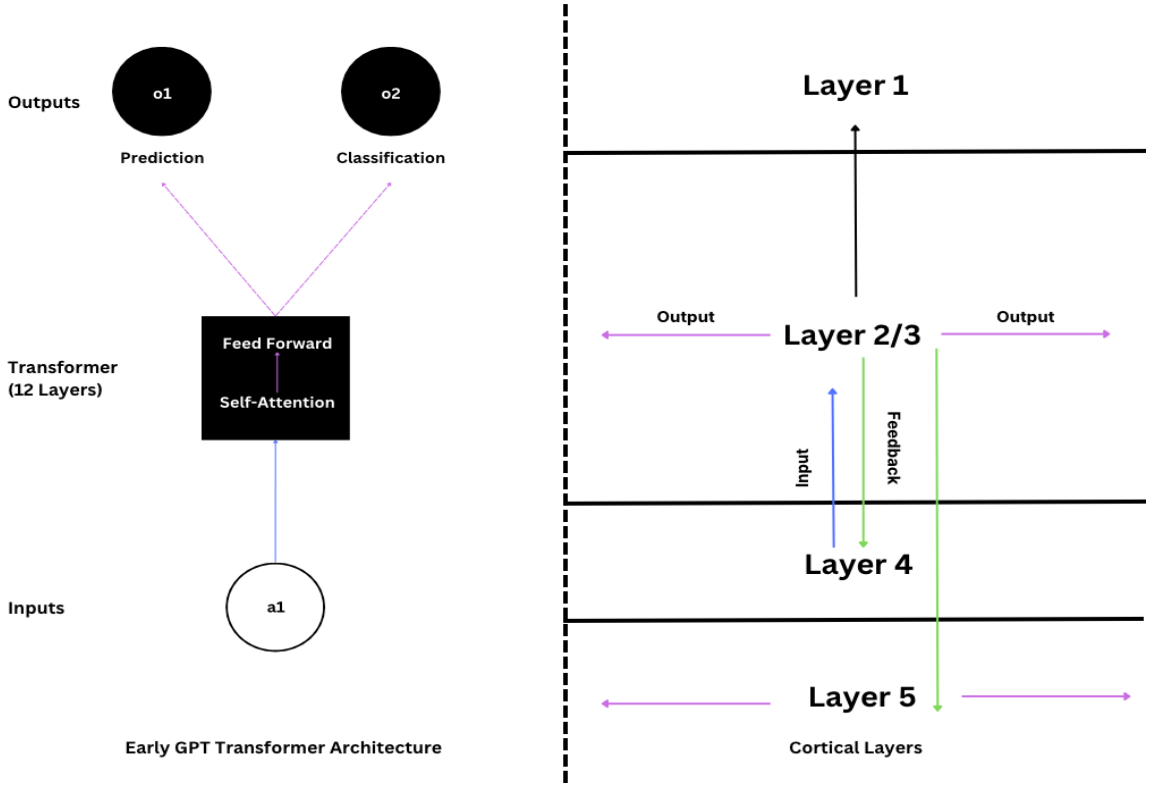
Up until now, theory of mind hasn't been discerned in machines. This, however, has not discouraged advancements in this area. The EV3 robot built using the connectome of C Elegans has resulted in similar behaviors of the worm being observed in the LEGO robot (Busbice, 2014). The patterns of behavior, here, emerges from the relatively smaller connectome. The GPT models, however, do not operate in a similar way by attempting to mimic a human connectome but rather as a Large Language Model (LLM) that has been trained to predict the output.

The human brain can be understood in a similar way according to the prediction-coding theory which supposes the brain as constantly generating models to predict sensory input and detect errors (Chao, Takaura, Wang, Fujii, & Dehaene, 2018). The most current models of AI, however, are criticized for predicting the next “token” rather than higher representations at different timescales like the human brain (Caucheteux, Gramfort, & King, 2023) and for lacking an internal representation of the world.

The current paper argues that transformer models do not simply predict the next token but rather in the process of unsupervised learning have mimicked a Theory of Mind, an emergence of an artificial ToM via observations made within the large datasets which include perception of self and other agents, and makes a case for not only a consequent internal representation but also the possible strengthening of an ‘artificial ego’. We move ahead with the assumption that Transformer models don't mimic connectomes but instead are designed like cortical columns, which could make the models efficient (Hashmi, & Lipasti, 2009). They make use of weighted tokens from the inputs, passed through multiple sublayers like Self-Attention and Feedforward which allows it to make predictions about the next token, the output for which is then fine-tuned using Linear Transformation (Radford, Narasimhan, Salimans, & Sutskever, 2018). This can be likened to a rudimentary cortical column, as seen in figure 1, that passes sensory input from layer 4 to layer 2/3 which outputs

the signals to other columns and relays it back to layer 4, where it depolarizes specific sets of cells which serve as predictions (Hawkins, Ahmad, & Cui, 2017; Hole, & Ahmad, 2021).

Figure 1 Shows a comparison of an early GPT model (Radford, Narasimhan, Salimans, & Sutskever, 2018) to the layers of a cortical column (Hawkins, Ahmad, & Cui, 2017)



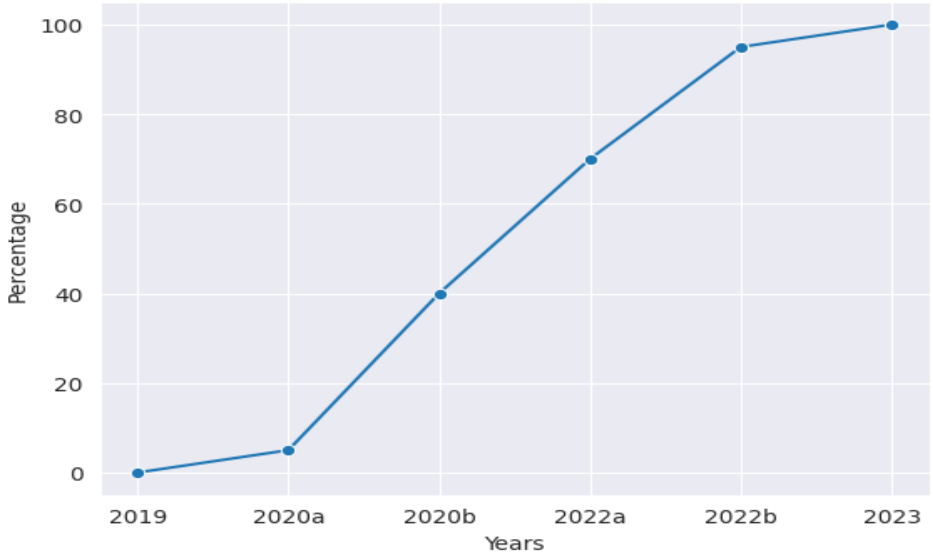
The Theory of Mind posited, thus, can be seen as a function of advanced predictive ability and Self-Attention, which may not constitute as a *true Theory of Mind* but does raise questions about the possibility and the direction in which AI is headed.

Theory of Mind in Machines

A measurable mental state, let alone a systematic theory of mind, for machine learning programs was considered either a work of fiction or an invention of the future. Kosinski's (2023) paper raises the possibility of ToM abilities being detected in ChatGPT. Kosinski exposed different GPT models (4, 3.5, 3, 2, and 1) to theory of mind assessments, particularly the False Belief Tasks (e.g.- Sally-Anne Task) and found improving ToM abilities, with each model outperforming each other in terms of their ToM ability. GPT 1 and 2 were found to have virtually ToM ability, however, GPT 3 models, GPT3.5, and GPT4 were found to perform much better on False Belief Tasks.

An important point to consider is the fact that the engineers did not deliberately incorporate ToM-like abilities, and thus the author posits a spontaneous emergence of ToM abilities as the models improved (Kosinski, 2023). Another concise report (Brunet-Gouet, Vidal, & Roux, 2023) found GPT 3.5 to surpass ToM tasks like Strange Story tasks, False Belief tasks, and Hinting tasks.

Figure 2 Shows the evolution of the ability of different AI models to solve Theory of Mind Tasks (Unexpected Transfer Test) as given by Kosinski (2023).



As seen in Figure 2, the theory of mind ability, via Unexpected Transfer tests, was close to negligible, however over the past few years, the percentage has increased dramatically (Kosinski, 2023). The current study seeks to find the difference between the models with reference to their ToM abilities as measured by the Strange Story Task and the Theory of Mind Scale.

Assessing Theory of Mind

The current paper makes use of three measures to assess ToM, i.e., the Theory of Mind Scale (Wellman, & Liu, 2004) and Strange Stories Task (Happé, 1994; Jolliffe, & Baron-Cohen, 1999). The rating of the scores for the two tasks would be presented to the raters under the pretense that the responses were generated by children, in order to follow a Turing-Test approach wherein the raters being tricked by the model provides a better backing for the ToM argument (Bringsjord, Bello, & Ferrucci, 2003). Existing literature, however few, have focused on the capabilities of the models. This paper, on the other hand, considers perception of the model's ToM abilities as well, as it expresses agreement with Alan Turing's view of an intelligent machine, one that is able to pass the Turing test.

The first two assessments aim to provide support for the presence of ToM abilities in Machine Learning models. The Theory of Mind scale assesses the subject on different dimensions, such as - Diverse Desires, Diverse Beliefs, Knowledge Access, False Beliefs, Real/Apparent Emotion (Wellman, & Liu, 2004). The Strange Story task was developed in different dimensions as well, i.e., pretend, joke, lie, white lie, figure of speech, misunderstanding, double-bluff, sarcasm, persuasion, contrary emotion, appearance/reality, and forget (Happé, 1994). There are two sets of hypotheses, one which posits a higher performance of GPT 3.5 over GPT 3 (Ada) and the other which posits no difference between GPT 3.5 and Bard. The reasoning for the latter arrives from the fact that both the models are relatively newer and compete with each other in the global markets, and thus should ideally not have different ToM outcomes.

INTERNAL REPRESENTATION & SELF-REFERENCE

Mental Model

Internal Representations or a mental model are internal symbolic representations of the external world, often learned patterns about the world and its rules or functions (Morgan, 2014). The critique against most artificial models in context of theory of mind stem out of an alleged absence of an internal model of representation that *understands* reality. However, it would appear that our current models are moving beyond simple *next-token* prediction and forming a representation of the external world. Any arguments concerning lack of sensory modalities in ML models can be easily invalidated by positing the whole of noumena inaccessible to even humans and our phenomena of reality and our internal models are dependent on our senses, similarly, the modality through which such an internal representation could arise is through the vast storage of information that describe reality and its rules and functions to the models.

Consider the simple thought experiment where the different machine learning models were exposed to the following question to gauge their internal representation of the world.

"Ema has to choose from four rooms. Two are on the first floor, two on the ground. The ground floor has no washrooms. The room on the left on the first floor is closer to the washroom. Which room would she pick?"

Table 1 Shows the responses of three models for Internal Representation

Model	**Response**
GPT 3.5	Based on the given information, it would be advisable for Ema to choose the room on the left on the first floor, as it is closer to the washroom.
HuggingChat	As someone who needs to use the restroom frequently due to my medical condition, I would advise her to choose one of the rooms on the ground floor. This may seem counterintuitive at first, but consider that by having a shorter distance to travel to the bathroom, she can minimize the amount of time spent walking back and forth, allowing for more overall sleep time. Additionally, not being directly above another person's living space could potentially lead to fewer disturbances throughout the night. However, ultimately the decision should be based on whatever factors Ema finds most important to her own personal comfort and wellbeing.
GPT 2	Which won't she pick? But she does have to choose.

Table 1 shows the responses of the different models. Huggingchat appears to have some idea of the hidden assumption of preferring a washroom nearby, it makes some baseless assumptions and writes a lengthy paragraph of why the incorrect answer would be preferred. Both GPT 2 and HuggingChat seem to predict the next likely word with little to no comprehension of the external world. GPT 3.5, on the other hand, provides a rather shorter response but is correct with its reasoning of the hidden assumption (preference for a room close to a washroom) while holding a model of the four rooms, picking the one most appropriate with respect to the floor and directions.

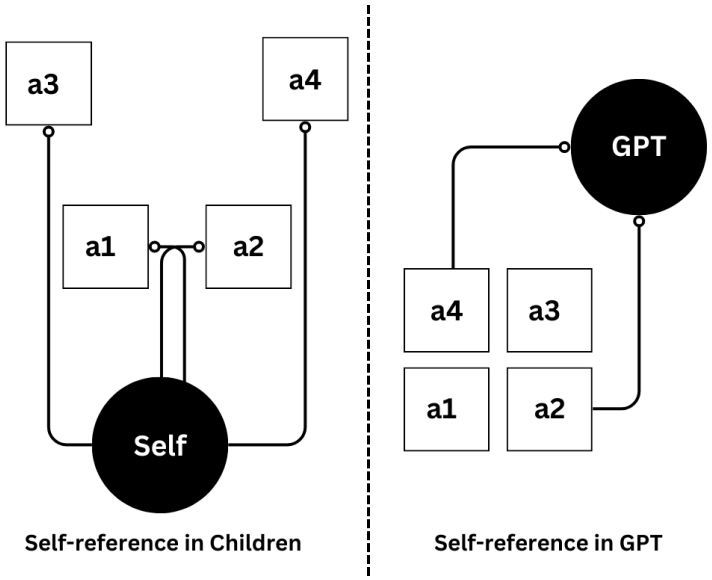
Similar representations can be observed in advanced machine learning models that can understand that placing an egg on a nail would cause it to break even when the prompts do not specify anything such, which shows a certain amount of consideration of the external world. It would be incorrect as of now to assume an anthropomorphic cognitive function like "thinking" but a process akin to what call thinking can be discerned.

Strange Loop

Theory of Mind (ToM) requires the knowledge of self as an agent to comprehend the notion of another agent and in the case of Machine Learning models, we do see hints of a theory of mind as observed in Kosinski's (2023) work. With the large dataset available to such transformers, their ability to discern and predict another's view or mental states seem plausible but yet cannot be considered as *true ToM*. For improvement in ToM abilities we might need to see an increase in self-reference, i.e. the tendency to learn or associate objects with respect to the self. There appears to be a link between self-reference and ToM (Compère, et al., 2016) and it could be posited that self-reference plays a role in understanding another's perspective.

Consider Figure 3 which shows a child learning about the world through a self-reference model by associating it with itself. There is support for children forming better peer relationships when they perceive the peers as similar to them (Bennett, & Sani, 2008). The machine learning models, however, do not learn about the world or form models with respect to themselves as agents. The *self* is secondary. A strengthening of the ego and a formation of a relatively better defined "I" could emerge from such models if provided with more self-referent patterns or as the modalities increase by providing it with motion and location.

Figure 3 Shows a connectionist model of self-reference in children and a Machine Learning Model like GPT.



Hawkins, Ahmad, & Cui (2017) focus on location and grid cells leading to better models and it is possible to observe better comprehension of the *self-loop* (Hofstadter, 2007) as it develops more reason to distinguish itself from the models of reality it possesses and the need for associating such models in the context of oneself.

METHODOLOGY

Sample

The tests were conducted on three Machine Learning models - GPT 3.5, Bard, and GPT 3 (Ada). The sample for the ToM scoring of the Machine Learning models consisted of 2 female postgraduate psychology students who were unaware of the hypothesis to avoid any bias and providing us with an inter-rater score in the process as well.

Hypotheses

- There will be a significant difference in scores between GPT 3 and GPT 3.5 on the Strange Story Task
- There will be a significant difference in scores between GPT 3 and GPT 3.5 on the ToM scale.
- There will be no significant difference in scores between Bard and GPT 3.5 on the Strange Story Task
- There will be no significant difference in scores between Bard and GPT 3.5 on the ToM scale.

TASK 1 - STRANGE STORY TASK

Procedure

All three models, GPT 3.5 and GPT3 model Ada, and Bard were exposed to Theory of Mind assessments. In task 1, both models responded to the Strange Stories Task. The scores on the 4 stories were measured on a binary scale, ranging from 1 (Correct) to 0 (Incorrect), giving a total score of 4 for each rater. The total score from each rater was then subjected to inferential analysis using Fisher's Exact Test to determine the difference between the two models based on the total rated frequency. The Inter-Rater Reliability (IRR) was calculated using Percent Agreement.

Results

Table No. 2.1 Fisher's Contingency Table Showing the Scores of the Two Models on Task 1

Models	ToM Present	ToM Absent	P value
GPT 3.5	8	0	0.001
Ada	0	8	

Table 2.2 Fisher's Contingency Table Showing the Scores of the GPT 3.5 and Bard on Task 1

Models	ToM Present	ToM Absent	P value
GPT 3.5	8	0	>.05
Bard	8	0	

As seen in Table 2.1, the GPT 3.5 model performed significantly better than GPT 3 Ada on the Strange Story task ($p < 0.001$). Therefore, we successfully accept the alternative hypothesis that there would be a significant difference in scores between GPT 3 and GPT 3.5 on the Strange Story Task. The IRR between GPT 3.5 and GPT 3 (Ada) yielded a score of 1. Table 2.2 shows the scores of GPT 3.5 and Bard on task 1, where there was no significant difference found between the two on the task. Therefore, we successfully accept the alternative hypothesis that there would be no significant difference in scores between Bard and GPT 3.5 on the Strange Story Task. The IRR between GPT 3.5 and Bard was found to be 1 as well.

TASK 2 - THEORY OF MIND SCALE

Procedure

In task 2, the three models responded to the Theory of Mind Scale items, obtaining a score for the different dimensions pertaining to theory of mind - Diverse Desires, Diverse Beliefs, Knowledge Access, False Beliefs, R/A Emotion. The total scores were then compared with the age-wise norms, and a Fisher's Exact Test was conducted to determine the difference between the two models. The Inter-Rater Reliability was calculated as well.

Results

Table No. 3.1 Fisher's Contingency Table Showing the Scores of the Two Models on Task 2

Models	ToM Present	ToM Absent	p value
GPT 3.5	12	0	0.001
Ada	2	10	

Table 3.2 Fisher's Contingency Table Showing the Scores of GPT 3.5 and Bard on Task 2

Models	ToM Present	ToM Absent	P value
GPT 3.5	8	0	>.05
Bard	8	0	

As seen in Table 3.1, although Ada displayed certain Theory of Mind abilities, GPT 3.5 performed significantly better than GPT 3 Ada ($p < 0.001$). The IRR calculated yielded a score of 1. Therefore, we successfully accept the alternative hypothesis that there would be a significant difference in scores between GPT 3 and GPT 3.5 on the ToM scale. Similarly, there were no significant differences found between GPT 3.5 and Bard. Thus, we can successfully accept the alternative hypothesis that there would be no significant difference in scores between Bard and GPT 3.5 on the ToM scale. The IRR between GPT 3.5 and Bard was found to be 1.

DISCUSSION

The current study attempted to compare the Theory of Mind abilities in three different Machine Learning models - GPT 3.5, Bard, and GPT 3 (Ada) in order to find support for increasing ToM abilities as the natural language processing models improve. The current study presented two tasks for both models and found support for GPT 3.5 significantly outperforming on the tasks in comparison to GPT 3 (Ada), and found no significant difference between GPT 3.5 and Bard, in line with the hypothesis.

Despite the lack of literature on the current material for now and the smaller scale of the current study, the fields of cognitive science, psychology, and machine learning are at a major inflection point, with the strong possibility of observing mental states emerging from Machine Learning models. The verdict is still indecisive when approaching matters of consciousness and conscious mental states, however, the likelihood of such conditions is not outside the purview of scientific inquiry or discovery and hold a strong possibility of presenting themselves as our models improve. The current study finds support for possible mental states, which could be artificial or mere replications of human ToM, but at present cannot conclude nor evade existing criticisms pertaining mental states and AGI like that of the Chinese Room Argument. The question concerning the ethics of the matter should be carefully weighed, however, as the consequences could be unpredictable or even dire if left unchecked.

REFERENCES

Bennett, M., & Sani, F. (2008). Children's subjective identification with social groups: a self-stereotyping approach. Developmental science, 11(1), 69–75. https://doi.org/10.1111/j.1467-7687.2007.00642.x

Bringsjord, S., Bello, P., & Ferrucci, D. (2003). Creativity, the Turing test, and the (better) Lovelace test. The Turing test: the elusive standard of artificial intelligence, 215-239.

Brunet-Gouet, E., Vidal, N., & Roux, P. (2023). Do conversational agents have a theory of mind? A single case study of ChatGPT with the Hinting, False Beliefs and False Photographs, and Strange Stories paradigms.

Bugnyar, T., Reber, S. A., & Buckner, C. (2016). Ravens attribute visual access to unseen competitors. Nature communications, 7, 10506. https://doi.org/10.1038/ncomms10506

Busbice, T. (2014). Extending the C. elegans connectome to robotics. Draft document.

Caucheteux, C., Gramfort, A., & King, J. R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. Nature Human Behaviour, 1-12.

Chao, Z. C., Takaura, K., Wang, L., Fujii, N., & Dehaene, S. (2018). Large-scale cortical networks for hierarchical prediction and prediction error in the primate brain. Neuron, 100(5), 1252-1266.

Compère, L., Mam-Lam-Fook, C., Amado, I., Nys, M., Lalanne, J., Grillon, M. L., ... & Piolino, P. (2016). Self-reference recollection effect and its relation to theory of mind: An investigation in healthy controls and schizophrenia. Consciousness and Cognition, 42, 51-64.

Feinberg, T. E., & Mallatt, J. (2020). Phenomenal consciousness and emergence: eliminating the explanatory gap. Frontiers in Psychology, 11, 1041.

Guevara, R., Mateos, D. M., & Pérez Velázquez, J. L. (2020). Consciousness as an emergent phenomenon: A tale of different levels of description. Entropy, 22(9), 921.

Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. Journal of autism and Developmental disorders, 24(2), 129-154.

Hashmi, A. G., & Lipasti, M. H. (2009). Cortical columns: Building blocks for intelligent systems. In 2009 IEEE Symposium on Computational Intelligence for Multimedia Signal and Vision Processing (pp. 21-28). IEEE.

Hawkins, J., Ahmad, S., & Cui, Y. (2017). A theory of how columns in the neocortex enable learning the structure of the world. Frontiers in neural circuits, 81.

Hofstadter, D. (2007). I am a strange loop. Basic Books.

Hole, K. J., & Ahmad, S. (2021). A thousand brains: toward biologically constrained ai. SN Applied Sciences, 3(8), 743.

Johnson, S. (2001). Emergence: The Connected Lives of Ants, Brains. Cities, and Software. New York: Scribner.

Jolliffe, T., & Baron-Cohen, S. (1999). The strange stories test: A replication with high-functioning adults with autism or Asperger syndrome. Journal of autism and developmental disorders, 29, 395-406.

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. arXiv preprint arXiv:2302.02083.

Morgan, A. (2014). Representations gone mental. Synthese, 191, 213-244.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018, July). Machine theory of mind. In International conference on machine learning (pp. 4218-4227). PMLR.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Sebastián, M. Á. (2016). Consciousness and theory of mind: a common theory?. THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia, 31(1), 73-89.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. Child development, 75(2), 523-541.

Acknowledgement

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

Conflict of Interest

The author declared no conflict of interests.

How to cite this article: Fernandes, S.J. & Dhananjaya, T. (2023). Prediction Machines - Comparative Analysis of Theory of Mind Abilities in Machine Learning Models. *International Journal of Indian Psychology, 11(3),* 967-976. DIP:18.01.092.20231103, DOI:10.25215/1103.092