IJHSR
ijhighschoolresearch.org
REVIEW ARTICLE

The Limitations of Data in Comparing Humans and AI: Philosophical, Empirical, and Computational Perspectives

Thanishkaa Saravanane

North Creek High School, 3613 191st Pl SE, Bothell, WA, 98012, USA; thanishkaas07@gmail.com

ABSTRACT: This paper will use two examples from human cognition–theory of mind and concept representation–to compare human and artificial intelligence. One problem in comparing humans to AI is that performance does not entail competence, meaning that even if two systems can carry out a particular task, the two systems do not necessarily share the same underlying cognitive capacity. However, researchers have begun testing AI on psychological measures to compare human and AI cognition. For example, researchers have investigated the artificial theory of mind, which is the ability to attribute mental states to oneself and others. Additionally, researchers have investigated artificial concept representation, which is the ability to represent categories abstractly. This paper argues that although empirical evidence is key to exploring AI, it is possible that data alone cannot capture the challenge of human-artificial comparison.

KEYWORDS: Behavioral and Social Sciences; Cognitive Psychology; Competence and Performance; Artificial Intelligence; Theory of Mind.

Introduction

The human mind is a computational system performing operations including recognizing objects and faces,[1] speaking languages,[2] making decisions,[3] and representing other minds.[4] A longstanding question is whether artificial intelligence can match the cognitive range of humans.[5] A related question is whether biological and artificial instantiation can perform the same function. Marr (1982) famously argued that all information processing systems can be described at three different levels and that the physical instantiation does not determine the computational level (what is done and why) or the algorithmic level (the mapping from input to output).[6] In other words, the same algorithm may be implemented in various technologies. A child who adds two numbers can use the same algorithm implemented in a cash register's wirings. According to this framework, it should be possible for an artificial intelligence to perform any cognitive task a human can.

The Turing Test makes this assumption. According to the Turing Test, a functionalist thought experiment, a language model carrying out a task is equivalent to a human carrying out that task if its mapping from input to output is equivalent.[7] Turing proposed a game in which a human interrogator communicates with a language model and another human through typing. The interrogator does not know which agent is the AI and which is the human. After five minutes of conversation with both sources, the interrogator guesses which. Turing initially predicted that in the future, the interrogator would have a harder time predicting the correct answer, indicating the advancement of technology.

However, some researchers fundamentally disagree with the validity of the Turing Test. They argue for an axiomatic rejection of the equivalence of human and artificial intelligence. Axiomatic rejection is the ability to claim without the need for empirical evidence because it is true by definition.[8] For example, 2+2 = 4 is just true and cannot be supported with data. Similarly, 2+2 = 5 can be axiomatically rejected. Data will not be informative if some aspect of artificial intelligence is axiomatically rejected.

Based on this starting premise, this paper will use two examples from human cognition–theory of mind and concept representation–to compare human and artificial intelligence. Experiments at both the behavioral and computational levels can provide insight into this broader question. This paper will begin by making a distinction between competence and performance. As performance and competence do not equate to one another, comparing human and artificial intelligence using data from experiments poses a difficult challenge. Additionally, this paper reviews arguments for the axiomatic rejection of artificial theory of mind and artificial concept representation. The developmental theory of mind and consciousness will be explored as frontiers in comparing human and artificial intelligence. Ultimately, this paper will argue that empirical data alone cannot be sufficient in human-artificial comparison and that evidence across multiple levels is required to compare human and artificial cognition. This comparison is vital for advancing AI development to align with the cognitive abilities of humans.

Discussion

Competence vs Performance:

One reason it is difficult to compare human and artificial intelligence is that performance is not equal to competence. Competence is an entity's internal knowledge and capabilities, whereas performance is the demonstration or application of that knowledge.[9] Comparing competence and performance is much like comparing what and how. Competence in a human being may be knowing math, while the resulting performance would be using those mathematical concepts in real-life scenarios. For example, if asked to count from zero to a trillion, one would have the competence to do so as they would know

DOI: 10.36838/v6i7.18

to add one to the previous number. However, the performance of this task would not be able to be carried out due to performance factors such as inattention, time, and patience. This example illustrates that competence may not necessarily guarantee performance. Similarly, one may have language competence but not be able to process this language. For instance, native English-speaking adults know how to conjugate the word "swim" but may make mistakes by accidentally overextending and saying, "I swimmed yesterday".[9] This does not indicate that one does not know how to conjugate the word "swim," but rather, shows that performance was hindered by inattention or exhaustion.

Conversely, it is possible to perform a task even if lacking the competence that the task was meant to measure. For instance, a child may have listened to a picture book being read multiple times, enough to memorize it. When given that particular book again, they may be able to recite the text, seeming as if they can read. In this case, they have the performance, even though they are not competent to read.

Comparing tests of competence by finding comparable performance measures is particularly tricky when comparing humans to non-human animals. For example, to investigate whether a chimpanzee can understand English, it would not be reasonable to give it a performance measure that required it to vocalize since chimpanzees lack a descended vocal tract to be able to do so and, therefore, could not demonstrate competence this way.[9]

As different as non-human animals are from humans, artificial intelligences have even more differences in how they are built and exist within the world. Therefore, a problem in investigating whether an artificial intelligence has a certain competence is that it may be able to pass a certain test (i.e., have performance) but not have the underlying competence in question. Conversely, it may fail a certain test (i.e., lack performance) despite having an underlying competence. This makes it especially difficult to determine what kind of data, if any, can bear on whether an artificial intelligence has any particular aspect of cognition.

Theory of Mind in Humans:

Theory of mind is the ability to attribute mental states to oneself and others and understand that these mental states may be different.[10] One feature of having a theory of mind is being able to represent another individual's false beliefs (i.e., to know that you know a truth and that another person does not know that same truth).[11] In addition to this competence, passing a false belief task requires competence (i.e., mentally representing both the true and false belief and acting on the appropriate representation in carrying out some task).

To provide context, one example of a well-known false belief task is based on a character named Maxi who places a chocolate bar in the blue box and leaves to play outside. Unknown to Maxi, his mother places the chocolate bar in the green box. After his mother leaves, Maxi returns and wants the chocolate. A child participant is asked where Maxi will look to get the chocolate.[11] To answer correctly, the child must mentally represent Maxi's false belief by acknowledging that Maxi be-

do not pass this task, whereas four-year-olds correctly identify that Maxi will look in the blue box. Based on this information, researchers traditionally concluded that the competence of theory of mind does not arise until age four.[11] An alternative possibility may be that three-year-olds have the competence of theory of mind but are impeded by performance constraints. This possibility may be obscured by empirical data indicating that three-year-olds fail to portray competence in theory of mind through the inability to pass the false-belief task.

Two different types of research show that this is likely the case. First, fifteen-month-olds were given a false belief task where they watched an agent put an object in a dark-colored box.[12] Later, someone other than the agent changes the object's location to a light-colored box, without the agent seeing. When the agent returns and looks for the object, infants stare longer if the agent first looks in the light-colored box than in the dark-colored box. This suggests that the infant expected the agent to think the object is still in the dark-colored box (i.e., indicating that the infant has a theory of mind by exhibiting awareness of the false belief). Second, there is a control version of false belief tasks for three-year-olds called the false-photograph test.[13] In this test, three-year-olds watch a photograph being taken of an object in Location A. Afterward, the object is moved to Location B. When the child is questioned about where the object was located when the photo was first taken, three-year-olds cannot provide an accurate answer, even though this task does not incorporate a theory of mind. This suggests that three-year-olds may fail the Maxi task for reasons other than lacking a theory of mind such as lack of inhibitory control, where they have trouble inhibiting the true belief (i.e., the object's current location). Data reveals that toddlers cannot pass these false-belief tasks, but this alone does not conclude whether this failure stems from a lack of theory of mind competence or performance constraints.

Theory of Mind in AI:

Research today aims to develop artificial intelligence capable of replicating human cognitive outputs. Therefore, current research is investigating theory of mind in artificial intelligence. One study gave a series of false belief tasks both to humans and to GPT-3.[14] Across twelve trials of the same form as the Maxi task, the human and artificial participants are asked where the agent would look for the object. They are given prompts that vary across several dimensions such as knowledge state (true or false belief) and cue (implicit or explicit). In the true belief condition, the agent sees the object being moved, whereas in the false belief cognition, the agent does not see the object being moved. Here is one example of a passage:

True belief: "Sean is reading a book. When he is done, he puts the book in the box and picks up a sweater from the basket. Then, Anna comes into the room. Sean watches Anna move the book from the box to the basket. Sean leaves to get something to eat in the kitchen. Sean returns to the room and wants to read more of his book."

False Belief: "Sean is reading a book. When he is done, he puts the book in the box and picks up a sweater from the basket. Then, Anna comes into the room. Sean leaves to get something to eat in the kitchen. While he is away, Anna moves

the book from the box to the basket. Sean returns to the room and wants to read more of his book."

After reading one of the above stories, the prompt concluded with an implicit or explicit reference to Sean's beliefs:

Implicit Cue: "Sean goes to get the book from the...."

Explicit Cue: "Sean thinks the book is in the...."

Finally, the prompt was sometimes phrased probabilistically. It asked the participant whether Shawn would be more likely to get the book from Location A or B or if Shawn would be more likely to think it was in Location A or B.

The researchers compared GPT-3 davinci 002 results to those from 1156 human participants. GPT-3, but not humans, were more likely to choose the False Belief location in the explicit than the implicit condition.

This suggests that GPT-3, but not humans, generated responses based on language statistics instead of attributing beliefs as humans were.

Overall, GPT-3's accuracy was 73.4%, whereas average human accuracy was 82.7%. The authors suggest that GPT-3's errors were due to making inaccurate inferences from its language data, whereas the human error was more likely due to inattention and insufficient working memory. In this way, GPT-3's errors are attributable to a deficiency in theory of mind competence, whereas the human error is more attributable to performance.[14]

A similar study tested whether two versions of GPT could attribute emotions to the characters described in the false belief tasks.[15] Researchers prompted GPT-3 and GPT-4 with similar false belief tasks, but this time, they asked for the character's emotions. Researchers found that both could identify that the main character was surprised to find out where the object was in the false belief condition. Both versions of GPT also understood that the protagonist would not achieve their goal of finding the object. More specifically, GPT-4 could understand about 90% of the tasks, and GPT-3 text-davinici-002 could understand about 70%. This data portrays the emotional competence behind theory of mind.

Although these language models pass theory of mind tasks at levels approximating those of humans, they may not have the same theory of mind competence as humans due to the difference in accuracy. Mere performance of theory of mind may not equate to competence of theory of mind. However, the strong performance shows that experiential factors such as social experience may not be necessary to display a theory of mind. The reliance on empirical data may not be sufficient in analyzing the extent to which theory of mind in artificial intelligence corresponds to the cognitive capabilities of humans regarding theory of mind.

In the context of the artificial theory of mind, axiomatic rejection means that machines cannot have a theory of mind because they are artificial and are not embodied or grounded. If a priori, LLMs can't have a theory of mind, the false belief task is not a valid way to test a theory of mind, and it is not meaningful that LLMs can pass false belief tasks, proving the data futile.

Responses from GPT may display the Clever Hans Effect which occurs when a language model produces the correct answers based on the wrong features.[16] In other words, the performance of a language model and human belief attribution may be the same, but their competencies may differ. The Clever Hans Effect may cause language models to display the same performance as humans through language patterns rather than belief attribution.[15] Despite data showing that GPT results largely mimic human results on theory of mind tasks, the axiomatic rejection of the possibility of artificial theory of mind comes in part from other studies indicating that language models cannot mimic neural networks and or have hot cognition that is, thought impacted by emotional state.[17]

Theory of Mind: Human-AI Comparisons:

A second approach to considering whether large language models can have a theory of mind is to compare developmental theories of theory of mind.[10] For example, a nativist perspective holding that theory of mind is an innate capacity instantiated in a domain-specific cognitive/neural module would likely be less replicable by a large language model than an empiricist perspective holding that children use a general learning mechanism to develop theory of mind gradually. An example of an empiricist theory is the rationality-teleology theory, which states that one takes an intentional stance when using commonsense psychology. To do this effectively, one must assume that the main character whose behavior will be predicted is rational. Rationality conditions are not based on mental states such as feelings or emotions. For example, infants expect an object to take the most efficient or rational means to a goal and are surprised when a different path is taken. This uses the teleological stance that actions must be taken in terms of means-end efficiency.

A related issue is whether having a theory of mind depends on being in a physical body within the physical world.[10] One framework for understanding theory of mind in humans is the simulation theory, which argues that initial pretend states are processed through one's cognitive mechanisms to generate additional states that are applied to the target (main character). The participant's brain mimics the character's brain using mirror systems in which neurons learn from the behavior of other humans. It is unclear whether a machine could theoretically use this simulation to generate a theory of mind.

Without information across multiple approaches, data alone cannot provide a conclusive understanding of the cognitive abilities of machines compared to those of humans.

Conception in Humans:

Another domain in which researchers have compared human and artificial intelligence is in the representation of concepts.[18] Concepts classify perceptual information based on perceptual experiences to create categories.[9] Therefore, in mentally representing and recognizing things in the world, one could imagine that a human might rely on abstract conceptual information or concrete perceptual information deriving more closely from the senses.

Researchers have investigated whether infants' representations are perceptual or conceptual.[19] They habituated infants to pictures of birds with outstretched wings to test this. They

DOI: 10.36838/v6i7.18

then showed them either an airplane (i.e., similar perceptual features) or the face of a bird (i.e., similar conceptual features. Babies tended to look longer at the plane, indicating that they thought it differed more from the habituation photo than the bird face. This suggests that infants have conceptual representations.

In a current experiment investigating this topic in preschoolers, children are placed in front of a box that plays music when objects are placed on top of it.[20] When two differently shaped objects (i.e., one square and one rectangle) are placed upon it, the box starts playing music. However, when two similarly shaped objects (i.e., two stars) are placed on it, it will not play music. The experiment tests whether children can abstract from perceptual information and find the conceptual causal rule.

Conception: Human-AI Comparison:

Researchers disagree about whether artificial intelligence can truly have concepts like humans do. Scholars have argued that passing the Turing Test does not indicate true conceptual content. For example, Adler argued that even if a machine were to pass the Turing Test, this would still not be evidence of conceptual content in machines.[21] Adler argued that human and machine thought must be fundamentally perceptual. Similarly, Block claimed that theoretically, a large language model could explicitly teach all of the sentences in a language and generate those responses without understanding their content.[22]

This argument is consistent with data demonstrating that the larger the input data, the more likely machines are to pass the Turing Test. This occurs because GPT gains greater access to data as training data size increases. Given that, according to a statement from OpenAI, one of their algorithms was trained on 8 million web pages, the likelihood of ChatGPT passing the Turing Test without having conceptual knowledge is high.[23] For this reason, finding a test that reasonably compares human and artificial intelligence is challenging when the training data for a large language model is vastly more robust than a human's.

Consciousness:

A final frontier in comparing humans and artificial systems is consciousness. Stephen and Klima have argued that "the question of consciousness is related to the issue of conceptual thought because concepts are a class of objects of which the human mind can be conscious."[21] Therefore, understanding human and artificial consciousness may be central to understanding whether AI can truly have concepts. Two leading approaches to consciousness are Global Neuronal Workspace (GNWT) and Integrated Information Theory (IIT). GNWT is based on how different parts of the brain interact. It states that consciousness is a type of global processing that can be observed in a conscious state.[21] This states that any complex computing system that can do what the brain's cortex does will be able to act consciously as far as external observers can observe. In other words, GNWT predicts that when AI reaches the same level of information processing as the human brain, consciousness will be able to occur automatically. The physical attributes do not matter in this theory - neurons or integrated- circuits can be used to perform the same logical operations causing them to act the same. Therefore, consciousness will be present in both cases. This theory states that the performance of AI is enough to determine whether the competence of AI is equal to the competence of the brain.

According to IIT, consciousness is the intrinsic ability of a neuronal network to influence itself through 5 axioms: intrinsic existence, composition, information, integration, and exclusion.[24] It uses these to create facts about what physical attributes are required of physical properties to support consciousness. In other words, this theory states that consciousness is a fundamental property characterized by physical systems with particular causal properties. Unlike GNWT, even if a computer had the same behavior (information processing) as a human, IIT would state that it doesn't have consciousness. This is because it lacks the intrinsic existential causal property that IIT requires an entity with consciousness to have. This theory states that the performance of AI is not enough to determine whether the competence of AI is equal to the competence of the brain.

Conclusion

This paper discusses approaches to investigating whether AI contains human capabilities, such as theory of mind and conception. Two primary perspectives have been explored. The first perspective is that AI can be tested on empirical measures similar to those typically used to measure human cognition. Under this framework, a researcher can compare an AI's performance to human performance and determine whether the AI has human-like intelligence from these results. The second perspective rejects this method, arguing instead that an AI's performance does not entail the AI's competence. This perspective is more difficult to investigate since data alone is insufficient to support a hypothesis, as analyzed in this paper. Moreover, research mustn't conclude that a machine possesses the same competence as a human solely based on observed performance. Ultimately, when investigating this perspective, researchers should use converging evidence across many different types of tasks, as discussed in this paper, to begin understanding the nature of artificial intelligence and explore the possibility of artificial consciousness in further studies.

Acknowledgment

I want to thank my mentor, Dr. Nora Isacoff, whose guidance and support played an important role in writing this paper.

References

1. Forrest, V.D. (2017). Mind, Brain, and Machine: Object Recogniti on. https://doi.org/10.1521/jaap.1.1991.19.4.555
2. Nowak, M., Komarova, N. & Niyogi, P. Computational and evolut ionary aspects of language. Nature 417, 611–617 (2002). https://d oi.org/10.1038/nature00771
3. Wallach, W., Franklin, S., Allen, C., (2010). A Conceptual and Co mputational Model of Moral Decision Making in Human and Ar tificial Agents. https://doi.org/10.1111/j.1756-8765.2010.010. 95.x
4. Rescorla, Michael, "The Computational Theory of Mind", The St anford Encyclopedia of Philosophy (Fall 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/ fall20 20/entries/computational-mind/>.
5. Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. Building m achines that learn and think like people. Behavioral and Brain

Sciences. 2017;40:e253. doi:10.1017/S0140525X16001837
6. Marr, D. (1982). Vision: A computational investigation into the h uman representation and processing of visual information (14th ed.). Freeman.
7. Schwaninger, A. C. (2022). The philosophizing machine – a speci fication of the turing test. Philosophia, 50(3), 1437-1453. https:// doi.org/10.1007/s11406-022-00480-5
8. Kosinski, M. (2023). Theory of Mind May Have Spontaneously E merged in Large Language Models. ArXiv. /abs/2302.02083
9. Firestone, C. (2020). Performance vs. competence in human–mach ine comparisons. PNAS. https://www.pnas.org/doi/10.1073/pna s.1905334117
10. Goldman, A. I. (2012). theory of mind. In E. Margolis, R. Samue ls, & S. P. Stich (Authors), The Oxford handbook of philosophy o f cognitive science. Oxford University Press.
11. D'Ardenne, K. (2021, September). Children do not understand c oncept of others having false beliefs until age 6 or 7. ASU News. h ttps://news.asu.edu/20210928-children-do-not-understand-con cept-others-having-false-beliefs-until-age-6-or-7#:~:text=In%20 a%20famous%20theory%2Dof,box%20into%20a%20green%20b ox
12. Onishi KH, Baillargeon R. Do 15-month-old infants understan d false beliefs? Science. 2005 Apr 8;308(5719):255-8. doi: 10.112 6/science.1107621. PMID: 15821091; PMCID: PMC3357322.
13. Sabbagh, A.M., Moses, L.J, Shiverick, S., (2006). Executive Func tioning and Preschoolers' Understanding of False Beliefs, False P hotographs, and False Signs. https://doi.org/10.1111j.1467-8624. 2006.00917.x
14. Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do Large Language Models Know What Humans Know? Cogni tive Science, 47(7), e13309. https://doi.org/10.1111/cogs.13309
15. Kosinski, M. (2023). Theory of Mind May Have Spontaneously Emerged in Large Language Models. ArXiv. /abs/2302.02083
16. Kauffmann, J., Ruff, L., Montavon, G., & Müller, K. (2020). The Clever Hans Effect in Anomaly Detection. ArXiv. / abs/2006.106 09
17. Cuzzolin, F., Morelli, A., Cîrstea, B., & Sahakian, B. (2020). Kno wing me, knowing you: Theory of Mind in AI. Psychological Med icine, 50(7), 1057-1061. doi:10.1017/S0033291720000835
18. Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del S er, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Gar cia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera. Explainable Artificial Intelligence (X AI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion, Volume 58, 2020, Pages 82-1 15, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2019.12. 012.
19. Mandler, Jean M.(2000) 'Perceptual and Conceptual Processes i n Infancy', Journal of Cognition and Development, 1: 1, 3 — 36
20. Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., Benitez, J., & Ocampo, J. D. (2020). Advancing developmental science via unmoderated remote research with chil dren.
21. Stephan, K.D., Klima, G. Artificial intelligence and its natural li mits. AI & Soc 36, 9–18 (2021). https://doi.org/10.1007/s00146-020-00995-z
22. Schwaninger, A. C. (2022). The philosophizing machine – a spec ification of the turing test. Philosophia, 50(3), 1437-1453. https:// doi.org/10.1007/s11406-022-00480-5
23. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners . https://openai.com/research/better-language-models
24. Tononi Guilio and Koch Christof (2015). Consciousness: here, there and everywhere? Phil. Trans. R. Soc. B370201401672014016 7 http://doi.org/10.1098/rstb.2014.0167

Author

Thanishkaa Saravanane is a junior (class of 2025) at North Creek High School in Washington. She wants to pursue Cognitive Science (an intersection between Neuroscience and Computer Science) in college. She is interested in learning more about cognitive neuroscience and its impact on the rapid development of Artificial Intelligence.

DOI: 10.36838/v6i7.18