Exploring Intelligence: Understanding Clever Hans and the Chinese Room

Lily Liang
EMPLID: 23938218
City College of New York
lliang002@citymail.cuny.edu

Abstract

This paper dives into an exploration of two significant discoveries, Clever Hans and the Chinese Room surrounding artificial intelligence. Clever Hans, the intelligent horse, not only illustrates the challenge of determining the extent of an animal's genuine comprehension but also compels us to reevaluate the boundaries of animal cognition and the intricacies of behavioral responses. In a similar situation, the conceptualization of the Chinese Room by John Searle triggers profound consideration of whether machines can genuinely grasp the essence of comprehension, thereby sparking curiosity about the limitations inherent in computer functionalities. Additionally, it delves into the criticisms and ongoing debates that revolve around these crucial topics. This paper closely examines these discoveries to understand the true extent of intelligence in both humans and machines. It seeks to unravel the complex dynamics that contribute to our understanding of intelligence and the role that computers play within this framework. These discoveries hold significant positions within the historical landscape of AI, showcasing the obstacles and breakthroughs that have shaped the evolution of intelligent systems. The paper conducts a comprehensive analysis of the impact of these milestones on our perception of AI, emphasizing their profound implications for the future trajectory of research in the field of artificial intelligence.

Introduction

The intersection of human cognition and artificial intelligence has long been a subject of profound fascination and inquiry. Over the years, there have been various milestones that have shaped the trajectory of AI research, revealing both the remarkable capabilities but also the limitations of intelligent systems. Among these remarks, two significant discoveries such as Clever Hans and the Chinese Room have stood out in helping us understand how complex the relationship between human intelligence and machine cognition can be.

Clever Hans, the remarkable horse that appeared to possess mathematical abilities, challenged the scientific community's understanding of animal intelligence and revealed the difficulties of unintentional cueing and the potential for misinterpretation in evaluating cognitive performances. The Chinese Room, a thought experiment proposed by John Searle, provocatively interrogated the philosophical underpinnings of machine intelligence, raising fundamental questions about the nature of consciousness, understanding, and the limits of computational systems.

This research paper dives into the historical context, implications, and significance of these case studies within the broader landscape of AI research. By examining the narratives and implications of Clever Hans and the Chinese Room, this paper aims to examine the relationship between human cognition and machine intelligence, and the evolving debates surrounding the nature of artificial and natural intelligence.

Clever Hans

1.1 Background and Context

Back in the early 1900s, there was this horse in Berlin called Hans that everyone talked about. People believed he could understand and answer questions by tapping his hoof to show letters or numbers. But guess what? Turns out, Hans wasn't some math genius horse after all. He was just picking up on tiny cues from the people asking the questions. This whole thing caused a big stir and made scientists realize they needed to be more careful when studying animal smarts which is a lesson we're still trying to stick to today.

This whole thing started way back in 1904 in Berlin, and it's what everyone calls the "Clever Hans Phenomenon." The horse could perform arithmetic operations using an alphabet with letters such as (A = 1, B = 2, C = 3 …), read the clock, recognize and identify playing cards, and know the calendar of the whole year.[5] To respond to any question, he would tap his hooves to indicate the numbers or even spell the name of the painter. However, it was later discovered that the horse was not performing these tasks through cognitive abilities. Instead, Clever Hans was responding to subtle cues from the questioners, unintentional movements, or changes in posture that indicated when he should stop tapping. The horse didn't learn to solve the problems by doing math or reading, it simply caught onto small hints from the experimenter's body language to know what answer to give.[4] This revelation revealed that the horse's abilities were not based on understanding, but rather on an ability to interpret human body language. As a result, it was discovered that the horse was an exceptional observer, able to discern minute facial cues from its trainer which allowed it to signal the right letter or number accurately, often just before or while tapping, in anticipation of receiving a reward.[5]

1.2 Significance in the History of AI

Even in today's advanced deep learning systems (and even basic regression), we face a similar challenge. One issue in machine learning is called the Clever Hans Effect. It's when the algorithm seems to have figured out a solution, but it's just catching on to random stuff in the data that doesn't matter.[3] For example, when something in the data seems closely linked to the right answer (you know, like the experimenter's body language), but it's not the real reason the answer's right (like doing the math correctly).[4] Even if the model is giving the correct answer, what if it was fooled easily due to relying on the wrong features? One example includes an algorithm distinguishing between wolves and huskies, but it mainly tells them apart by the presence of snow in the image which is a disaster for any serious application.[4] The thing is, it's hard to tell if your network is making these types of mistakes because they don't show up in the classification error since you need a dataset without any biases or a really good test set. However, in any dataset, there is always bias, especially in huge datasets, it’s harder to find. Machine learning algorithms can sometimes exhibit a phenomenon akin to the Clever Hans effect, wherein their apparent intelligence stems from capturing patterns within the data that lack relevance to the underlying problem they aim to address.[3] It’s important to be more aware of

this effect because it can lead to inaccurate or biased predictions especially when it pertains to algorithms that impact our daily lives. For example, imagine this credit score algorithm learns from past data, but the data has biases against certain groups, like minorities or folks with less money. So, the algorithm ends up being biased without even knowing it. We can also think about a medical diagnosis tool that learns from images of tumors but it might not be spotting cancer itself but just the way the images were taken.[3] By understanding the Clever Hans Effect, data scientists and ML developers can develop more robust and accurate machine learning systems that truly learn the patterns in the data that are relevant to the problem at hand and this process would involve careful feature selection and regularization to avoid overfitting, rigorous testing and evaluation.[3] Just because Clever Hans got it wrong, it doesn't mean that we should be taking that risk, especially in high-stakes applications such as healthcare and finance that result in real-world consequences. We can use this effect as a lesson to dive into what the data truly means and what needs to be analyzed carefully. Thus, the lesson we learned is that we have to be careful to reduce its impact when creating machine learning systems.

Chinese Room

2.1 Overview of the Thought Experiment

The Chinese Room was presented by philosopher John Searle in 1980, where the Chinese Room argument challenges the idea that machines can ever truly be intelligent and the concept of strong artificial intelligence. He proposed the idea that a person who does not understand Chinese can simulate understanding by following a set of rules. He imagines himself in a room with boxes of Chinese characters he can't understand and a book of instructions, which he can. The person is given a set of symbols and rules for manipulating them and is then asked to translate Chinese sentences into English which means this person can follow the rules and produce correct translations, even though they do not understand the meaning of the Chinese sentences.[1] If a Chinese speaker outside the room passes him messages under the door, Searle can follow instructions from the book to select an appropriate response. The person on the other side would think they're chatting with a Chinese speaker - just one who doesn't get out much.

2.2 Criticisms, Debates, and Philosophical Implications

Now, according to Alan Turing, the father of computer science, if a computer program can convince a human they are communicating with another human, then it could be said to think. The Chinese Room suggests that, however well you program a computer it doesn't understand Chinese, it only simulates that knowledge which isn't intelligence. But as humans, we are trained that way ever since going to school so that also means humans aren't that intelligent either. The fundamental question it raises is what does it mean to "understand" something and can a system that follows rules without comprehending be considered intelligent?[1] Searle's saying that just following rules isn't understanding. He thinks the person in the room is just doing some technical stuff with symbols, not getting what they mean. That's why he believes computers, which work only based on rules, can't understand anything. For example, the input is the person inside who doesn't recognize questions in Chinese but somehow, by following the program's instructions, this person in the room manages to send out Chinese symbols that are correct answers to these questions(the output). So, the program makes it seem like the person in the room understands Chinese, but in reality, they don't understand a single word of it. According to Searle, "Here's the thing: if the guy in the room doesn't get Chinese just by following the right program, then no other digital computer does either. That's because no computer, as a computer, has anything that the guy doesn't have."[1]

Other critics said that his point relies on a lot of assumptions like understanding language needs you to have some sort of personal experience or awareness. However, some philosophers and AI experts reckon it might be doable to make machines that handle language without needing those personal experiences.[1] Another assumption includes that the person in the room is only following a set of rules but some argue that the person is engaging in a form of interpretation. They're saying the machines can pick the right rules depending on what the Chinese sentences mean which is a type of understanding, even if it is not conscious or subjective.[1] Another issue with Searle's argument is that it draws a hard line between how words are structured and what they mean, which might not match how humans use language. Searle argued that computers don't understand meaning the same way as humans can but the point is much more than just that. It's saying that meaning isn't only about moving words around, and that you need things like what's going on around you, your own experiences, and what you know about your culture to get language. This goes against how Searle's argument is challenged since he is making the idea of

understanding language too simple while his experiment assumes that language is just about symbols and rules, but it's way more complex than that.

Language isn't something you can just move around like that cause it's always changing and tied to our experiences and stuff. One example could be Chatbot GPT-3, like the Chinese Room in Searle's argument, it processes information based on syntactical rules and patterns without really understanding. Despite its remarkable ability to generate coherent and contextually relevant responses, GPT-3 lacks inherent comprehension, as it operates solely on processing input data and generating output based on learned patterns. This goes along with what Searle's saying, that machines can't grasp language or ideas 'cause they don't have real thinking abilities or awareness. Even though GPT-3 seems to understand with its fancy language skills, it's not getting what it's saying as we do. We know that language is shaped by so many factors that play a role in shaping the meaning of our words so his experiment doesn't accurately represent the way that humans use language, and therefore cannot be used to conclude the limitations of machines.[1] It's also worth noting that he didn't intend this to be an argument against AI but he used the example in a border argument against the idea that mental states and processes can be defined purely in terms of their functional roles, rather than their physical properties so he argued that functionalism can't provide a complete account of human cognition.[1] The Chinese Room proved that a system can carry out complex tasks without actually understanding what it is doing.

2.3 Importance to AI

Searle's argument challenges the idea that you can make a computer truly understand things so computers don't understand what the words mean in a deeper sense, they just move symbols around based on rules. During his time, it was the early days of AI so they relied on a very outdated view of what machines can and can't do but as time passed, AI became so advanced that they could perform complex tasks like recognizing images or speech without relying on rules or instructions.[1] This shows that his argument was probably the foundation of what AI can do better, we can create machines that exhibit intelligence and understanding but they still don't work the same way as our brains. Now this escalated fairly to the point where researchers challenged each other in public debates on the topic of whether machines can ever be truly intelligent but during this time, the field of AI was rapidly advancing.

Many thought they could create machines that could exceed human intelligence. If we look at Google's virtual assistant, it can make phone calls and book appointments for people without them realizing they are speaking to an AI. To a certain point, we can't distinguish between an actual person versus a machine but have they succeeded in becoming more intelligent than us? Engineers are working behind the scenes to make this AI we use every day so intelligence is more than having the right answers as Jeff Hawkins writes: "We are intelligent not because we can do one thing particularly well, but because we can learn to do practically anything."[2] Searle would have agreed with Hawkins and he is the reason why such advancements came in the next few decades. The main point is understanding how humans learn, and think and the cognitive interaction between our bodies and the context surrounding us.[2] Another interesting point is that when IBM's Deep Blue won against Garry Kasparov in 1997, the Chess grandmaster said: "Anything we can do (…) machines will do it better (…) If we can codify it, and pass it to computers, they will do it better."[2] Kasparov is saying that even though we're trying to make computers do what we do, we still don't know how our brains work and how we learn languages. There's still a lot we don't get about how we think. So, is it possible that we can't make super smart computers because we don't understand our brains well enough?

Conclusion

In conclusion, Clever Hans and the Chinese Room have contributed significantly to our understanding of the intricate relationship between human cognition and artificial intelligence. Through Clever Hans, we learned how smart animals can be but how hard it is to assess cognitive abilities, particularly in non-human entities. This study shows us that our surroundings in experiments are important and must be considered before making any conclusions. Meanwhile, the Chinese Room allows us to examine underlying meaning with machine intelligence, raising questions about consciousness, comprehension, and the potential limitations of computational systems. It got us thinking about whether machines truly understand things making us question how smart they are. As we keep learning about AI from these remarkable discoveries, these important cases show us how tricky intelligence is and how it affects both computers and humans. Knowing these things helps us make smarter AI systems that understand how complex intelligence can be and how tough it is to make machines as smart as people.

Works Cited

1. Alogu, Sud. “The Chinese Room Problem. Exploring The Limits of Machine… | by Sud Alogu | Medium.” *Sud Alogu*, 12 March 2023, https://sudalo.medium.com/the-chinese-room-problem-71eeda00deb. Accessed 4 November 2023.
2. Barrero, Andres Felipe. “Can AI Think? Searle's Chinese Room Thought Experiment.” *TheCollector*, 16 February 2023, https://www.thecollector.com/can-ai-think-searle-chinese-room-argument/. Accessed 4 November 2023.
3. Chettri, Bhusan. “The Clever Hans Effect in Machine Learning: an overview by Bhusan Chettri - IssueWire.” *Issuewire*, 24 July 2023, https://www.issuewire.com/the-clever-hans-effect-in-machine-learning-an-overview-by-bhusan-chettri-1767401061368276. Accessed 4 November 2023.
4. Lindwurm, Eugen. “Deep Learning, Meet Clever Hans. An article about hidden mistakes made… | by Eugen Lindwurm.” *Towards Data Science*, 15 August 2020, https://towardsdatascience.com/deep-learning-meet-clever-hans-3576144dc5a9. Accessed 4 November 2023.
5. Samhita, Laasya, and Hans J Gross. “The “Clever Hans Phenomenon” revisited - PMC.” *NCBI*, 13 November 2013, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3921203/. Accessed 4 November 2023.