Searle's Chinese Room Argument and Functionalism

Rohan Aslam

PHL242

Dr. Joshua Brecka

If you and a machine react the same way to an incident, does your reaction hold more meaning because you are a human with a mind and corresponding mental states? According to functionalism, it does not. The theory states that what makes something a true mental state is the role it plays, not what it is made of. (Levin, 2023) But if both a human and a computer system have the same reaction or mental state, does that also mean they have the same level of understanding? John Searle, in "Minds, Brains and Science", presents his Chinese Room thought experiment that builds the case to challenge this by showcasing that such systems lack true understanding. Within this argumentative essay, I will utilize the Systems Reply to argue that, although Searle's Chinese Room example presents a strong challenge to functionalism, it ultimately falls short by focusing solely on the person inside the room, who is only part of the system rather than the system as a whole.

Functionalism argues that mental states such as pain and desire depend on the function they have in a system, regardless of what that system is made of. (Levin, 2023) For example, pain is a mental state that can be caused through injury or harm, can cause anxiety and can bring about screaming because of the pain and the desire to deal with the injury to stop the pain. (Brecka, 2025, Slide 22) In humans, this chain of mental states may be caused by neurons firing in the brain. But functionalists argue that if such mental states and flows can occur in computers with no neurons or biological tissue, then these beings can be in pain too. (Levin, 2023) John Searle's Chinese Room argument is famous for looking to tackle this exact train of thought. Searle argues that systems that may be functionally correct and provide the correct answer still lack true understanding. (Searle, 1984, p. 33) To demonstrate his point, Searle proposes the Chinese Room thought experiment. Imagine that a person is locked in a

room with a set of rules for each Chinese character. But the person has no understanding of the language. Imagine that Chinese characters are passed into the room, the person uses the given rules to build the response and then passes it back out. The rules and work are set in such a way that each output is perfect, and what a native speaker would produce. Searle argues that although to someone outside the room it may seem that the person inside knows Chinese, that is just not the case. The person is simply following the rules and instructions given and has no understanding or knowledge about the characters themselves. So if simply going through the appropriate states of understanding Chinese is not enough to allow you to understand it, then neither is it enough for any digital computer. (Searle, 1984, p. 33) When I first heard this argument, it made sense to me. The person in the room is following predefined rules similar to a computer executing code corresponding to a task it's given. Neither seems to understand what they are doing, or understand the semantics behind the letters or words, but rather just follow the steps. However, after analyzing the argument more, I came to the objection that although the person does not understand Chinese, one could argue that the entire system of the room and the person do. This is known as the Systems Reply and provides an objection to Searle's argument.

The Systems Reply argues that the whole system understands Chinese. Although the reply acknowledges that the person inside the room manipulating symbols does not understand Chinese, that is only the processing part of the system. The baskets in the room, the rule book, the person and anything present to help with translation as a whole work together as a system which does understands Chinese. (Searle, 1984, p. 34) Hearing this argument initially, I was quite drawn to it. It made sense to me, and I found it relatable to human experiences. Humans have neurons in our brain that fire, which then cause chemical reactions leading to alteration of mental states and understanding. The neurons themselves, like the person in the room, may not understand everything, but the entire system, the brain, does. As a rebuttal to the Systems Reply, Searle argues that there is no way for the system to get to semantics from only the syntax, and the person in the room has no way of knowing what the

symbols mean, and neither does the system. (Searle, 1984, p. 34) From my understanding of functionalism, it only argues that if a system can transition from the correct state to the next, then it has understanding. In the Chinese Room example, characters are passed in, translated, and put in the appropriate bucket as output. The system goes from state to state and produces the correct output. I believe a human translator would go through the same flow in their mind when translating. The only addition would be their semantics to the words, which could also just be a similar system that produces a feeling from a word. An example of semantics in action can be seen in "The Matrix" (1999). The machines create a digital world where humans appear to have understanding and go through the mental states of pain, pleasure, etc. I would argue that if machines are unable to have understanding or semantics, do the humans in the digital world also lack that? I argue that systems like those in "The Matrix" can have understanding because they go through the exact mental states and processes they would in reality. Based on this thinking, I side with the Systems Reply as opposed to Searle's argument.

In conclusion, the Chinese Room thought experiment by Searle provides a strong challenge to Functionalism. However, it fails to consider the system as a whole. Based on the definition of functionalism and the Systems Reply, I found it a more convincing argument as I was able to relate the concept more closely to my own experiences and thinking. Based on the definition provided, if a system can go from one correct state to another, then, according to functionalism, it has understanding.

References

Searle, J. R. (1984). Can computers think? In *Minds, brains, and science* (pp. 28–38). Harvard University Press.

Levin, J. (2023). Functionalism. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2023 ed.). Metaphysics Research Lab, Stanford University.

Brecka, J. (2025). *Artificial intelligence* [Lecture slides]. University of Toronto.