Can the Machine Understand:
An Evidence-Based Approach to the Chinese Room

Keiland W. Cooper
Faculty Mentor: Dr. Frederick F. Schmitt, Department of Philosophy, *Indiana University Bloomington*

ABSTRACT

The debate of a thinking machine continues on, especially in an era where machines are achieving tasks that we never thought possible. In this essay, I explore one of the most famous critiques of the thinking machine, Searle's Chinese room, by breaking down his argument into two claims of varying scope. I then offer an alternative method to assess this argument by employing a top down approach, in contrast to Searle's method, which seems to advance from the bottom up. I explore the current thinking on how the human brain may come to understand the world, as well as some of the features of these semantics. This is all in an effort to elucidate some of the features necessary for machine understanding and to accurately assess whether a machine possesses them. I conclude that Searle may have been too quick to judge the abilities of computers and that a claim that any digital computer cannot understand is much too strong.

KEYWORDS: neuroscience, understanding, artificial intelligence, semantics

INTRODUCTION

The idea of a non-human machine that could perceive and understand the world as we do has enthralled humanity since the idea's conception. While the feat has yet to be accomplished, debates for its feasibility, especially with current standards of technology, remain bilateral. Few arguments against this strong Artificial Intelligence (AI), as it is called, are as famous and thought provoking as the Chinese Room argument posited by Searle in his 1980 paper, *Minds, Brains, and Programs* (Searle, 1980), which argues that no programmed digital computer can understand in the same way that humans do. In this discussion, I will paint a picture of a how we may come to our feat of understanding and explore some important properties of this trait: a worthwhile endeavor in its own right, but also in an effort to determine if Searle's broad sweeping conclusion is valid in light of evidence from cognitive psychology and neuroscience.

The Chinese Room argument was written to disprove the functionalist claims of AI by stating that even *if* a computer could exhibit human-like behavior; that the behavior is not a sufficient criterion for the computer to *understand* that behavior. For example, if the computer was to hold a conversation with a participant, as in the Turing test, tricking an observer to believe the computer is a human (Turing, 1950) or, as Searle cites, in Roger Shank's program, which had the ability to answer reading comprehension style questions about stories (Schank, 2013). These behaviors would be meaningless to the computer, and no understanding of its actions would commence. To illustrate this argument, Searle envisions himself being tested on his ability to hold a conversation by a third party. He communicates with this person by written Chinese symbols, where the pages are slid under the door into a closed room in which Searle sits. Searle is meant to communicate back to them in Chinese as well; however, a problem emerges as he does not know the language himself. Luckily, in front of him in the room there is a ledger, a large book which contains a proper reply to any question or statement written in these Chinese symbols that is or could be passed to Searle. All he must do then is look up the question-and-reply pair in the ledger, write that reply down, and then slide it back under the door. He could repeat this for the entire conversation with the person testing him. For the thought experiment it is assumed that the book is rich enough that Searle can produce a reply to any statement in the same way that a native Chinese speaker could if they were in the room, meaning that the person which Searle is communicating with would not be aware that Searle does not speak or understand Chinese. To the person sliding the pages under the door, it would seem that there is a native speaker behind it, and they would not know that Searle is oblivious to any of the meaning or content of the conversation.

Searle argues that his thought experiment is analogous to a seemingly intelligent computer or computer program and that his argument shows that while a computer program may appear to understand, as he did in the room, this may only be an illusion and the computer cannot understand. He goes on to say that the method of symbol lookup and manipulation as done in the room with the ledger is an apt description of how computers operate; their programming acts as the ledger does, and all they must do is match an input to a predetermined output.

It seems that in his argument two claims can be extrapolated regarding a programmed computer understanding. A weak version of the argument is that a computer will lack the ability to understand if it is only manipulating symbols as defined by its programming, as it cannot manipulate, or know the meaning behind them. As Searle would put it, the computer would have only a syntax but no semantics. To him, it would seem that the computer is empty, full of meaningless, arbitrary symbols while the true understanding is present only in the programmer. Beyond this, a stronger claim is made. He not only states that a computer will not understand if it only manipulating symbols, but rather that this lack of an ability is at the very heart of what a computer is: that *all* programmed digital computers can *only* manipulate symbols and thus do not contain the capacity for meaning. Because of this, all computers cannot understand or produce strong AI by the very virtue of being a computer.

Seemingly, Searle captures something with the weak claim that cannot be denied. A computer that is programmed to reply with a given sentence isn't any different than if you repeated a Chinese sentence that you were just told out loud, these would be as arbitrary of sounds to you as a gibberish made-up word. To put it another way, just because an agent *can* produce an action does not mean that the agent understands it. This, however, does not give reason to extrapolate Searle's stronger claim, where his reasoning invokes a bottom up approach, starting with the properties of the machine and applying them to a view of understanding. Searle begins with the computer's mechanics, and from there assumes this machinery

is incapable of a behavioral feat, the problem being that this line of reasoning does not address the qualities of understanding and meaning that cannot be captured by a computer. Moreover, without this adequate description of human understanding and the brain mechanisms therein, one is subject to overgeneralization, as I suspect Searle has fell victim to. Because of this, I view the problem inversely to his approach, proceeding with two guiding goals:

(1) Illuminate what mechanisms the brain may leverage to achieve its ability to understand

(2) Determine if the answer to (1) can be instantiated by a digital computer

If the answer to (2) is yes, then the stronger claim must be false, however, it would seem that Searle's argument has proceed directly to answering (2) in a somewhat hastily fashion, without first contemplating (1). I should clarify that the goal of the following sections is not necessarily to construct a robust framework for understanding and its implementation, but rather to judge if what we currently know of the phenomena of understanding, particularly from cognitive psychology and neuroscience, is sufficient to claim that a computer lacks the ample properties to achieve it.

HOW THE BRAIN UNDERSTANDS: LANGUAGE LEARNING

To address (1) we must explore the only known system which can understand the world to the best degree known, the human brain. While the exact mechanisms of this understanding are still a mystery, what is known so far is sufficient to offer insights into the current exploration of whether a computer could achieve a similar feat. While there is still debate on the degree to which innate knowledge plays a role in learning and understanding, many theories of learning rest under the assumption that higher level cognition is derived from experience (Griffiths, 2002; National Research Council, 2000; Bjorklund, 2017). This sensory level experience may also extend into higher level cognition (Barsalou, 2008). But, if understanding is learned, what is it that the brain is learning?

A canonical case to help demonstrate the brain's learning mechanisms is language learning, an especially apt example in this case as it is the modality most leveraged in the Chinese room argument. Language learning is a complex process spanning multiple developmental stages as well as many faculties (Bjorkland and Causey, 2017). While native arguments such as universal grammar have been poised and are still under much debate (Yang, 2004), it would seem that much of language is derived from experience (Yee, 2017). This can occur very early with children learning the language of which they are raised and exposed to: A child raised in China would learn Chinese. Of course, this makes sense; few would suspect that a child raised in a Chinese household only speaking Chinese would learn English, due to the lack of exposure to the language. Studies have validated this as well. Children who are not exposed to a language within the first few months of their lives will fail to acquire the ability to make that language's specific sounds, such as the /r/ sound in Japanese or the /r/ sound in English, which are quite distinct to the one who has learned each of the languages (Kuhl, 2005).

Within that environment, experience continues to drive language learning. In the earliest stages of a child's life, language learning requires a physical referent. Countless studies have reached the same conclusion (for review, see Roy, 2005). This makes sense, as one would not expect a child to babble about abstract ideas like "irony" or social institutions like "manager". Most, if not all, of a child's word learning at the earliest stages involve referents such as the baby's mother or father, or objects such as noodles or dog (Woodward and Markman, 1998). While the baby may fail to enunciate correctly, the ability to make a sound in association with an object is, without a doubt, apparent. This is also shown by very young children's inability to learn language from audio tapes or television shows indicating human interaction is required for them to successfully learn (Krcmar, 2007). It isn't until the baby is two or three that hints of abstract contents come to fruition. It is also around this time when the baby begins to refer to feelings or their inner state using words like "now" or "sad"(Bergelson and Swingley, 2013).

This background on language learning begins to paint a story of content. In the earliest stages, word learning demands a physical referent, a baby using the statistics of the environment can learn which sounds correlate to which objects, accelerated by the input from the mother or father. As the baby learns, it is able to understand abstracted content, and then over time, he or she can understand content from pure symbols alone, such as learning from text while reading. This begins to hint at a syntax from semantics argument, putting forward that semantics does not in fact come from syntax, but rather, in the earliest stages at least, semantics derives from experience. Later, once the child has a certain degree of knowledge, he or she can then learn from the symbols, as these relate back to the foundation of experience the child has acquired.

SEMANTICS ARE DYNAMIC

As learned from our exploration of language learning, understanding and meaning are ever changing as a function of experience. In terms of language used above, as we learn more and more about the world, we continuously gain more and more knowledge of a word's referent, resulting in a richer interpretation of meaning and deeper understanding. This process is what creates a distinction between a word's *intrinsic* meaning, the meaning agreed upon by multiple people as would occur in a dictionary and a word's *internal* meaning, how one interprets the meaning given their own experiences (Aitchison, 2012).

One way to view this is that a person's understanding could be more or less impoverished than another's. For example, a biologist who has studied rabbits for 15 years will have a richer internal understanding than a child who learned the word "rabbit" using one photo in a book. Both the scientist and the child will have the same referent when using the word rabbit, however, the amount of understanding, if quantifiable, would substantially differ. The scientist would have copious amounts of knowledge of the animal: she would know its evolutionary history, where and how they live, who their predators are, the rabbit she had as a child which got her interested in them in the first place, how they behave. This is in contrast to the child, where the only information the child would know of the animal would be only what was apparent in the photo. It should be noted that this is still a fair amount of information of the animal. For example, the child would know that the animal has fur, ears, eyes, a nose, is small, etc., but in relation to the information of the scientist, the child would have a smaller amount of knowledge.

It should also be noted that this view, while blatantly not the whole story, a correlation appears between the amount of similar, connected knowledge and understanding. The importance of this feature and the dynamic nature of meaning as a whole, can be illustrated colloquially as follows: a child who memorized the answers to his math test would be said to have a lesser understanding than the child who derived them from first principles. A major difference in this case, among many others, is that the child who derived the examples for himself has more experience and thus, more opportunities to gain knowledge

and information about the subject or problems he is attempting to solve and understand, while the child who memorized the answers still knows something about the problems, it is undoubtedly less than the other child would know. This idea can be transferred to language as well. If someone looked up "philosopher" in the dictionary, having no exposure to a philosopher or philosophy, they would have a narrower view than a graduate student in philosophy, having more experience with the word and all of its uses and meanings. While the memorized definition would tell the person something about the philosopher, it does not encapsulate the larger understanding that would come with experience. This is a largely relevant point for computer understanding, which will be illuminated in a later section.

SYMBOLS AND THE BRAIN

A point that has yet to be captured is how the brain could maintain and represent all of this learned information, or more specifically, how could the semantics that it learns be represented. As previously noted, while the precise answer is beyond the reach of current science, fundamentals can be derived, and hints do appear in the literature.

Primarily, the brain is thought to work with symbolic representations, not the most uncontroversial stance to hold (Lycan, 2015), and should be safe to assume for the purposes of this argument. The only information that can reach the brain is the activity from its sensory inputs. At the very beginning, these inputs convert into neuronal signals. These signals are what carry the information about the world and, thus, are the foundation for all other activities the brain is involved in. These are essential, as a brain without its inputs and outputs would be nothing but a lump of fat on a table. Moreover, the necessity of these sensory inputs has been well known to neuroscience for decades (Greenough, 1987). It could be that these sensory inputs, at the very beginning of a child's life, serve as the grounding of all other representations. Over time, the original representations could be moved to memory, allowing them to be altered and manipulated and to interact with the new incoming sensory information. Over time this process leads to an aggregate of information in the form of many edited representations of the world. This entire process is symbolic, yet from it, meaning can be derived.

CAN A MACHINE UNDERSTAND?

The preceding sections function to serve as a brief summarization of and exploration into the phenomena of the brain's understanding. The first was used to paint a story of a canonical case of how the brain learns the meaning in language, heavily leveraging experience and using what was learned in the past to understand and learn from the present. The second was used to demonstrate that this meaning can, and most likely will, change over time, and the third to show that the brain deals in symbolic representations and offers to offer insight on their dynamics. With these points in tow, we can then proceed to the second question poised in the beginning of this essay: could a digital computer provide the framework necessary to support understanding as a human would, using experience and the dynamic aggregation and nature of knowledge?

Primarily, the computer, like the brain, would need to have the means to interact with the outside world in some way. A brain that is removed from one's head without any of its sensory inputs and placed on a table will not learn, as no information or activity would be able to interact with it. The same could be said for a computer. This aspect is handled in Searle's argument, as a reply to his Chinese room, aptly named the Robot reply. It states that attaching sensors as inputs to his Chinese room and allowing it to interact with the environment would be a case for a computer to understand. For example, a video camera as input and/or legs to move, all in an effort to allow the computer to perceive the world. Searle noted that this changes nothing, as computer is in fact still a computer, collapsing the argument back into the Chinese Room. To him, you're only changing where the symbols come from, but not how they are manipulated. He simply put the Chinese room inside the robot instead of the computer, reducing the argument back into the original case.

The Robot reply is a good start, but I must agree with Searle's reply in this formulation in a sense that simply adding input from the world is sufficient to escape the Chinese room, though most likely essential if we wish the computer to learn to any degree. As noted above, a brain on a table without any input or output is quite unimpressive by all human standards of cognitive ability, thus the computer too would need sensory modalities of a sort for input and output, just as we ourselves do. Though in terms of understanding, Searle is right that these sensors are not enough, thus, it would seem that there is something still missing from the Chinese room that would allow it to understand when faced with the Robot reply.

Returning to our exploration of how the brain understands, we are reminded that a trend throughout was that the brain was ever changing: new information was coming in, and at all times, the brain was learning and leveraging what it had already knew, or in other words, the brain is highly dynamic. This contrasts with a major premise in the Chinese room with the ledger of Chinese symbols as it assumes that the room is static in how it can interact with the third person attempting to communicate with it and the outside world as a whole. This idea of such a static computer, as it bleeds through into Searle's argument of the room, hints at some of the larger issues with his interpretation of the computer as a whole, that it is only a box that can relay information as it is programmed or can only do what it is "told", so to speak. This may not have been far off from the time the paper was written, but he missed the potential these machines can achieve. In his view, the computer can only be the child that memorized the answers to the math exam, but not the one that derived them. I should note that I am not claiming that the computers can rise above their programming, at least not yet, nor that they have intentional states, but rather that they can be *programmed* to learn. Computers, at least by today's standards, are not limited to a list of *if-this-then-do-this* commands, but rather have the ability to change based on experience. Computers can learn from experience, and they can do so without any explicit programming of the task. Such programs have been able to teach themselves to play Atari video games (Mnih, 2013) or the ancient game of GO from scratch (Silver, 2017), each reaching superhuman abilities in breathtaking time. While it probably shouldn't be definitively concluded that these programs understand, it does illuminate that these machines can do more than just follow the if-then programming of which Searle's argument seems to allude, and it demands an updated, wider definition of what a computer is and can do.

Lastly, the larger problem of syntax and semantics can be addressed. As we have learned in earlier sections, meaning is primarily experience, as words refer to tangible objects in the first stages of our lives. With time we are able to move away from requiring a physical referent to having the ability to form more abstract associations using past experiences. Thus, the question for the computer's ability to understand is shifted from meaning and instead to experience. A new criteria for understanding may then be that if a computer can accurately represent what it would experience, could this experience then be manipulated as it learns? An objector may say no, that even if meaning is grounded in experience, a computer would be unable to represent it, and that the symbols of the computer are not apt for this ability. To this I say why? There is no evidence to support that a

symbol system *cannot* support experience, in fact the opposite seems true, as stated in the above section on symbols and the brain, for at the lowest level the brain takes in information in the form of sensory representations. While I will be the first to admit that the specifics of this process are currently beyond reach, an alternative theory would have to accommodate this fact.

Thus, to answer the second question of whether a computer is able to enact the important features of understanding: if a computer could interact with the world, have the ability to dynamically alter its states and the type of states it could be in, and have the ability to support the experience that it gained, it would seem that the strong claim of the Chinese room would be disproved. This, of course, is not an exhaustive, definitive list, but rather enough to show that the argument posed by Searle is too stringent. Though not inconsistent with the computer, these features would drastically alter the Chinese room. In this view, the room would not be a static book and operator, but rather an ever-changing enterprise, with groups of pages linked together and moved in large groups as more rush in. All sorts of processes would be occurring, as symbols would be associated, forgotten, copied, or altered. While the precise mechanisms of this enterprise are still out of reach, we widen our eyes to more of what's possible, in an effort to not only create the intelligent machine, but rather more importantly to understand the understanding of the creatures which wish to build it: ourselves.

AUTHOR INFORMATION

All correspondence should be sent to the first author:
keiland.cooper@gmail.com

REFERENCES

Aitchison, J. (2012). Words in the mind: *An introduction to the mental lexicon*. Hoboken, NJ: John Wiley & Sons.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol., 59*, 617-645.

Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition, 127*(3), 391-397.

Bjorklund, D. F., & Causey, K. B. (2017). Children's thinking: Cognitive development and individual differences. *SAGE Publications*.

Greenough, W. T., Black, J. E., & Wallace, C. S. (1987). Experience and brain development. *Child development*, 539-559.

Griffiths, P. E. (2002). What is innateness?. *The Monist, 85*(1), 70-85.

Krcmar, M., Grela, B., & Lin, K. (2007). Can toddlers learn vocabulary from television? An experimental approach. *Media Psychology, 10*(1), 41-63.

Kuhl, P. K., Conboy, B. T., Padden, D., Nelson, T., & Pruitt, J. (2005). Early speech perception and later language development: Implications for the" critical period". *Language Learning and Development, 1*(3-4), 237-264.

Lycan, W., (2015). Representational Theories of Consciousness, *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2015/entries/consciousness-representational/>.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with deep reinforcement learning. arXiv preprint *arXiv:1312.5602*.

National Research Council. (2000). How people learn: Brain, mind, experience, and school: Expanded edition. *National Academies Press*.

Roy, D. (2005). Grounding words in perception and action: computational insights. *Trends in cognitive sciences, 9*(8), 389-396.

Schank, R. C., & Abelson, R. P. (2013). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. London: Psychology Press.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences, 3*(3), 417-424.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Chen, Y. (2017). Mastering the game of go without human knowledge. *Nature, 550*(7676), 354.

Turing, Alan M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433-460.

Woodward, A. L., & Markman, E. M. (1998). Early word learning. *Handbook of child psychology: Vol. 2. Cognition, perception, and language,* 371-420.

Yang, C. D. (2004). Universal Grammar, statistics or both?. *Trends in cognitive sciences, 8*(10), 451-456.

Yee, E. (2017). Fluid semantics: Semantic knowledge is experience-based and dynamic. *The Speech Processing Lexicon: Neurocognitive and Behavioural Approaches, 22*, 236.