Proceedings of the 2nd International Conference on Interdisciplinary Humanities and Communication Studies
DOI: 10.54254/2753-7048/27/20231195

Rethinking the Chinese Room Argument

Feiruo Zhang[1,a,*]

[1]Department of management, Beijing Film Academy, Xitucheng Road, Beijing, China
a. csmithj51511@student.napavalley.edu
**corresponding author*

Abstract: With the development of artificial intelligence (AI), the long-lasting question was raised again in public debates, “Could a machine think?” When it comes to this profound issue, the famous Chinese Room Argument (CRA) invented by American philosopher John Searle must be considered. Since Searle first proposed the CRA in 1980, numerous scholars have discussed it over the past decades. However, there are still some problems regarding CRA such as why CRA is so powerful that it can always cause people’s confusion; could people use CRA to argue against nowadays artificial intelligence? These are the problems which will be discussed in this article. By reviewing Searle’s philosophical views and diverse replies to the CRA, the article will draw the conclusions that CRA is powerful because of intentionality, a key notion of Searle’s philosophy, and the CRA cannot be applied to the AI model which is widely used by most of nowadays AI, connectionism.

Keywords: the Chinese Room Argument, artificial intelligence, connectionism

1. Introduction

Artificial intelligence (AI) has emerged as one of the most debated and intriguing technologies in contemporary society. AI is able to carry out numerous kinds of job such as distinguishing human faces, drawing pictures and distributing advertisements. In some industries, the invasion of AI has caused protests and strikes. It is fair enough to say that AI has become an irreplaceable part of our life. In addition, the invention of ChatGPT, an AI program that can answer any question people ask about, might cause a lot of people to think that long-lasting question explored by many philosophers, scientific fictionists and directors, “Could a machine think?” And among all the theoretical discussions about this old but interesting question, probably the most famous and the easiest to understand is the Chinese Room Argument (CRA) invented by American philosopher John Searle in 1980. In CRA, Searle designed a thought experiment, using metaphors that is comprehensible for almost everyone, to draw the conclusion that machine, representing by digital computer running certain programs, do not have the ability to generate what people called minds or thoughts. Numerous scholars have discussed CRA from all kinds of view, some of them tried to overturn it and some supported it. But only a few of these articles have really explored why CRA seems so plausible, what makes it so powerful? Fewer of them have directly located and attacked the core premise of CRA.

This article will first introduce the Chinese Room Argument and some replies to it. And then in section 2, the article will explain why CRA is so powerful and analyse its theoretical resources. In section 3, the article will introduce two main schools of artificial intelligence in 3.1. After that, the article will demonstrate two refutations against CRA which support connectionism in 3.2. In the end, section 4 will be the conclusion of the article.

2. The Chinese Room Argument and Replies Against It

The Chinese Room Argument, proposed by the American philosopher John Searle, challenges the notion of “strong artificial intelligence” [1]. “Strong AI” proponents claim that human intelligence basically works like a computer, mind is its software and brain is its hardware. Moreover, they think a Turing machine, with the ability of manipulating and computing symbols under certain rules, is able to generate human intelligence. Opposing to such view, Searle conceived the famous Chinese Room Argument. In CRA, Searle envisaged that he is locked in a room alone with some cards which have Chinese characters on them and a book full of rules written in English. Searle in the room understands English but knows nothing about Chinese. People from the outside could hand him some cards with Chinese characters on them and then Searle has to choose some cards as a response. He is able to do that by checking that book full of rules. These rules are pure syntax instead of semantics. For instance, one of the rules might tell him to hand over the card with a symbol looks like an “x” with a short line above it when he receives a card that has a symbol looks like a “k” with a dot on its left. Assuming people who design the book are proficient in both Chinese and English. Therefore, people outside the room can always receive the right answers of the questions they send into that room. However, Searle in the room still does not understand Chinese at all. He does not understand the questions he receives and the answers he chooses. All he has to do is simply follow those purely syntactical rules. The situation of Searle in the room is the same of a digital computer, both of them do not understand the symbols they are dealing with even though they could always offer the right answers. Thus, proponents of “strong AI” are wrong about human intelligence since human are able to learn and understand symbols. Searle concluded the logical structure of Chinese Room Argument in *Minds, Brains and science*.

There are four premises:

(1) “Brains cause minds.”

(2) “Syntax is not sufficient for semantics.”

(3) “Computer programs are entirely defined by their formal, or syntactical, structure.”

(4) “Minds have mental contents; specially, they have semantic contents.” [1]

Searle draw his conclusion from premise (2), (3) and (4), “No computer program by itself is sufficient to give a system a mind. Programs, in short, are not minds, and they are not by themselves sufficient for having minds.” [1]

The Chinese Room Argument has caused a lot of refutations such as The System Reply, The Robot Reply, The Other Minds Reply and The Intuition Reply.[2] Proponents of The System Reply concede that Searle in the room does not understand Chinese. However, Searle in the room is just a part of the entire system, like a CPU in a digital computer. The entire system including syntactical rules, cards with Chinese symbols on them and Searle in the room does understand Chinese. Searle proposed a simple response to The System Reply. By remembering all the cards and the rules, Searle in the room can be the entire system himself. But since he still only follows purely syntactical rules to generate answers, Searle as the entire system also does not understand Chinese. The Robot Reply also thinks Searle in the room is unable to understand Chinese and therefore cannot generate human intelligence since people understand meanings of different symbols by interacting with them in the real world. For example, people understand the meaning of “book” because they have seen them in libraries or book shops, or read them, or heard of them. Therefore proponents of The Robot Reply claim that a computer with a robot body, arms, legs, camera and microphones and so on, can learn and understand Chinese. Searle fought back easily. He reckons all these arms, legs and sensors can do is just adding additional inputs, that is more cards with Chinese symbols. But the computer still processes these inputs in a purely syntactical way. Therefore, nothing significant has changed, the Robot still does not understand Chinese. The Other Minds Reply claims that people tend to use different standards when it comes to whether a computer has consciousness or not. For instance, people can easily know that an individual has intelligence and consciousness since they are able to communicate and interact with each other. But when it comes to computer, people always try to find more evidence than just behaviors. The famous Turing Test was designed for this, people should use the same standard of deciding whether other people are intelligent and conscious or not as deciding whether a computer is intelligent and conscious or not. Searle's response is simple, he thinks that we need something else except for external behavior to attribute intelligence and consciousness to something like a digital computer. For Searle, that is some internal states of minds such as intentionality. The standpoint of Searle is so called Internalism, contrary to Externalism which is held by proponents of The Other Mind Reply. The article will introduce Internalism and Externalism later in section 2. In the end, The Intuition Reply reckons the CRA is highly depended on people's intuition about consciousness, understanding and intelligence. In other words, Searle did not develop a precise definition of understanding and depended on people's intuition of understanding. However, The Intuition Reply might not be that fair as it seems. Searle holds an Internalism view of understanding, that means the main feature of understanding and human intelligence is intentionality, a mental state that refers to something. The article will demonstrate this technical term later in section 2.

3. The Core of Chinese Room Argument

The CRA has been widely discussed for decades since Searle first invented it in 1980. What makes CRA so fascinating that many scholars had been attracted to it? If CRA is really that easy to be overturned, like what section 1 has discussed above, why it seems so plausible in the first place? The answer is Searle's premise (2), syntax is not sufficient for semantics. This premise is the most crucial proposition in CRA. After an introduction of some terminologies, syntax, semantics and intentionality, the article will demonstrate two reasons for why premise (2) is the most powerful and important proposition in CRA. In linguistics, syntax and semantics are two separate subdisciplines. "Syntax means the study of set of rules governing the way that morphemes, words, clauses and phrases are used to form sentences in any given language." [3] As it can be seen from above, the description of syntax is very similar to that book full of rules in CRA. In other words, syntax is formal in the sense of it only deal with physical form of symbols instead of their meanings. In contrary, semantics is the study of meaning. It explores questions such as "are the meanings of words subjective or objective?", "what determines the meanings of words and sentences?" and so on. In addition, Searle thinks that semantics is highly connected with another terminology, intentionality.

Contemporary discussions of intentionality were launched by Franz Brentano and then developed by his student Edmund Husserl. In general, "intentionality is the power of minds and mental states to be about, to represent, or to stand for, things, properties and states of affairs".[4] But what is the relationship between semantics and intentionality? The answer to this question could be dated back to Searle's mentor John Langshaw Austin's theory of speech acts. The theory of speech acts is a part of a larger view of pragmatics, which holds that meanings of propositions or words should not be simply deduced from their truth value or logical form. Instead, people should focus on how these words and sentences are being used in different contexts. The usage of language is able to disclose the puzzle of meaning. Austin claims that there are three kinds of speech act, locutionary act, illocutionary act and perlocutionary act.[5] The attempts of reducing natural language to logical forms, leading by Frege and his truth value theory, can only be applied to a small part of natural language, locutionary act. But when it comes to, for instance, illocutionary act, the meaning of a sentence could not be defined by its truth condition. Sentence such as "I will come back tomorrow" is neither true or false, but happy and unhappy. By saying this sentence, people are trying to do something such as making a commitment instead of stating a fact. The structure of such speech acts could be written in the following form: “F(p)”. “F” stands for the power of illocutionary act and “p” stands for the propositional content. Based on Austin’s theory of speech acts, Searle developed his theory of intentionality. In his book *Intentionality, an essay in the philosophy of mind*, Searle said that “Intentionality is that property of many mental states and events by which they are directed at or about or of objects and states of affairs in the world.”[6] Besides, Searle specified the relationship between intentional states and speech acts, “Intentional states represent objects and states of affairs in the same sense of ‘represent’ that speech acts represent objects and states of affairs.”[6] In other words, Searle replaced “F” in “F(p)” by intentionality and symbolized it as “S(r)”. Therefore it is reasonable to conclude that in Searle’s view, semantics, using words and symbols meaningfully, is basically the same as intentionality, representing words and symbols meaningfully.

The first reason for premise (2) to be so essential is Searle’s responses to all the replies against CRA. Searle refuted almost all of those replies by depending on premise 2. For The System Reply, Searle said that “there is no way that the system can get from syntax to the semantics” [1] since the CPU, Searle in the room, has no idea what do these symbols stand for. For The Robot Reply, Searle also argued that such causal interactions between the robot and the real world simply just add some inputs of the system, it does not change the fact that syntax cannot generate semantics. Moreover, Searle stated that no matter what kinds of new technology or program are invented, a computer program cannot think, understand or be conscious. And the reason for this assertion is still the premise (2), syntax is not sufficient for semantics.

However, it is normal for people to question this premise. What makes it so powerful, is it just people’s intuition of understanding or it does have a solid philosophical foundation? The answer to such questions is the second reason for premise (2) to be so crucial. The article will explain why by summarizing philosophical discussions about intentionality. In the philosophy of mind, there is a debate between the Externalism and Internalism. Externalism holds the view that minds are not determined entirely by something inside people, including their brain. Therefore, Searle should be considered as a proponent of Internalism because of his premise 1, “Brains cause minds”. The article will not interfere this controversial debate between Externalism and Internalism. Instead, it will simply demonstrate this standpoint, intentionality cannot be reduced to some external affairs such as behavior. The reason for this is not complicated. People do not have to be a philosopher to know that what they have in mind sometimes differs a lot from what they actually do. For example, a man sitting in a room says “The door is open”. This same sentence could have different meanings depending on his mental states or in Searle’s words, intentionality. It could mean “Close the door please”, if the man feels the outside is too cold; It could mean “please leave”, if the man wants to end an unpleasant conversation with someone. Intentionality, the key terminology in Husserl’s phenomenology, highly accords with everyone’s daily experience and intuition, that is when we utter words or sentences, we represent, refer to or think of something which might influences meaning of our utterance and behaviors. Some Externalism such as behaviorism and functionalism believe that everything in mind could be reduced to something observable such as behaviors and computational programs. These methods are useful in some occasions, they can be explanatory. But in the case of CRA, intentionality can not be reduced or ignored because if critics remove intentionality from CRA, discussions of this topic will become meaningless. In other words, CRA is fascinating and attractive precisely because the view of intentionality and semantics, which highly accords with people’s intuition, is contrary to what takes place in artificial intelligence. All the replies to CRA discussed in section 1 did not focus on premise (2) and that is why even though people could easily understand these replies, the CRA can still cause their confusion. That means questions like how does syntactical program utter words and sentences meaningfully is still puzzling even if people understand The System Reply, The Robot

Reply or The Other Minds Reply and so on. In summary, to eliminate the confusion caused by CRA, people have to directly fight against the premise (2).

4. Connectionism Responses to CRA

4.1. Introduction of Symbolism and Connectionism

Section 3 delves into two predominant schools of thought in artificial intelligence and cognitive studies: symbolism and connectionism. Moreover, the article will show how proponents of connectionism argue against the premise (2).

The main principles of symbolism could be summarized as follows: “There are such things as symbols, which can be combined into larger symbolic structures. These symbolic structures have a combinatorial semantics whereby what a symbolic structure represents is a function of what the parts represent, and at the same time all cognitive (reasoning) are manipulations of these symbolic structures.” [7] In general, symbolism claims that minds should be modeled on the level of symbol. That is to say symbolic AI models follow certain rules to compute some given symbols. Computation in symbolic AI models only means to compute on the level of symbol, that is to relocate, combine or duplicate those given independent symbols, instead of learning, generating or disintegrating symbols. As can be seen from above, symbolism is susceptible to the critics from CRA since symbolic AI models only carry out syntactical missions which highly resemble the situation depicted in CRA.

On the other hand, connectionists reckon that a lower level could work better than the level of symbol, such as the level of neuron or the subsymbolic level. Basically, the network of a connectionist system is made of multiple layers of numerous units or so called neural nodes.

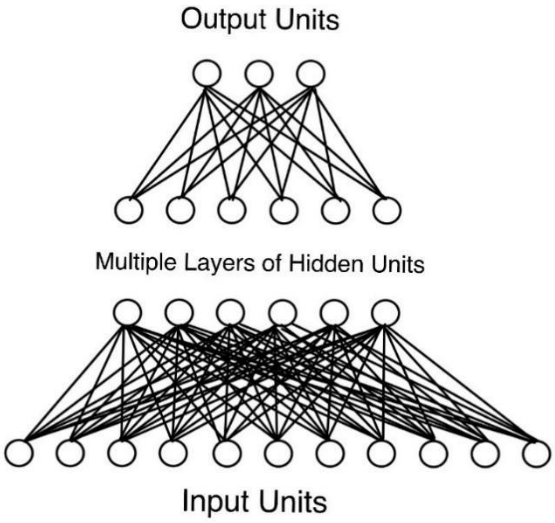

Figure 1: Illustration of connectionism network.

The function of input units is to generate certain value that represents some stimuli outside the system. As the illustration shows, these input units then send their activation value to the hidden units with which they are connected and these values will be sent to output units eventually. In addition, these values are calculated while being transmitted through different layers according to some functions and weights, which represent the strength of connections between units.[8] As the article has introduced above, connectionism AI does not work on the level of symbol. In other words, connectionist AI models do not directly compute some given symbols. Therefore, it makes CRA less persuasive by undermining the premise (2). These units in a connectionism network do not work like the man sitting in the Chinese room, they do not deal with given, discreet symbols. Instead, they act more like neurons in human's brain.

4.2. Two Connectionism Refutations Against CRA

4.2.1. The Luminous Room

The first refutation is The Luminous Room Argument, invented by the Churchlands in 1990. In their article "*Could a Machine Think?*", the Churchlands aimed at the premise (2), they called it "the crucial third axiom" [9], which is "Syntax by itself is neither constitutive of nor sufficient for semantics." The Churchlands said that the premise (2) is not true because it is not a logical truth but an empirical rule. For example, people in the 18th-century thought it is impossible for compression waves in the air to become sound. The Churchlands believe that the relationship between syntax and semantics is the same as waves in the air and the sound. They invented a thought experiment with the same structure of argument as CRA, The Luminous Room. In 1864, Maxwell said that the essence of light and electromagnetic wave is identical but this hypothesis was not accepted by many scientists at that time. Assuming someone who tried to refute Maxwell's hypothesis proposed an experiment as follows. A man stands in a dark room with a magnetic stick in his hands and he swings it up and down. According to Maxwell, this movement should generate light. However, the evidence of the experiment and people's intuition are contradictory to the hypothesis. Could this experiment be the falsification of Maxwell's hypothesis? The Churchlands proposed four premises by imitating Searle's structure of CRA.

Premise (1): Electricity and magnetism are forces.

Premise (2): The essential property of light is luminance.

Premise (3): Forces by themselves are neither constitutive of nor sufficient for luminance.

Conclusion: Electricity and magnetism are neither constitutive of nor sufficient for light.

Back in 1860s, most people would find this argument really plausible. But people now know that the waving frequency of the magnetic stick has to be above a certain level to make luminance, 1015Hz, which is far too high for manpower. Therefore, the premise (3) of The Luminous Room Argument is not true even though it accords with people's intuition and seems like a logical truth since force and luminance have completely different definition. The Churchlands claimed that Searle made the same fallacy in CRA. The premise (2) of CRA could be overturned by new scientific discoveries. That means it is an empirical statement instead of a "conceptual truth" like what Searle claimed it is. This critic to CRA by the Churchlands is reasonable and rigorous. Searle's premise (2) is not necessarily true because it is not logical truth and Searle provide no empirical evidence to prove it. Specifically, premise (2) is not logical truth, or analytic truth in Kant's words. The terminology Searle used to describe premise (2) is "conceptual truth" but he did not explain it further more. Literally, "conceptual truth" should have the same meaning as logical truth or analytic truth since analytic truth stands for propositions whose truth only depend on its definitions and concepts. For example, "A red flower is red" is a typical analytic truth. However, premise (2) should not be counted as analytic truth. Because the predicate, not sufficient for semantics, cannot be deduced from the definition of the subject, syntax. On the other hand, as an empirical statement, Searle did not provide any evidence to prove the premise (2). After refuting CRA, the Churchlands continued to argue whether classical AI, that is symbolism AI, can generate minds or not. Surprisingly, the Churchlands hold the same view regarding symbolism AI as Searle, they both think classical AI is not the solution of minds and cognitive science. But their reasons for this view differed. The Churchlands thought symbolism AI cannot produce a machine that could think not because "syntax is not sufficient for semantics", but because of a lot of empirical evidence such as symbolic AI models cannot deal with some specific tasks like learning;

and with the development of neuroscience, scholars found out the structure of symbolic AI differs from how brains actually work. Therefore, they supported another school of AI, connectionism.

4.2.2. Refutation to The Chinese Gym

Searle's response to connectionism is The Chinese Gym.[10] The Chinese Gym shares the same content and structure of CRA except that there are a lot of people instead of one man sitting in the room. Everyone in the gym is doing syntactical jobs, collecting inputs, calculating them by fixed weights and function and in the end send them to someone else. Therefore, Searle insisted that connectionism still cannot generate minds since these units of neural network are still purely syntactical. Besides The Luminous Room discussed above, this section will develop another way to argue with Searle.

In his book *Minds, Brains and Science*, Searle proposed his theory of the mind-body problem. In general, Searle reckons that there should not be dualism between mind and body, just like there is no such problems like "digestion-stomach problem". Firstly, the mind-body problem means question like: How does mental, immaterial things interact with physical, spacial things? To solve the long-lasting mind-body problem, Searle proposed two arguments. (1): All mental phenomena...are caused by processing going on in the brain. (2): All mental phenomena are features of the brain.[1] The problem of these two arguments is that how can minds be caused by the brains and at the same time be features of the brains. Searle provided his solution of the mind-body problem by answering this question. Traditional view of causal relationship thinks a causal event consists of two discreet events, one as cause and the other one as effect. Such conventional view makes people to accept some kind of dualism of mind-body problem since what happens in minds and brains are regarded as two set of discreet events. Materialism reduces everything takes place in mind to brain and idealism reduces everything in brain to mind. To avoid such dualism, Searle proposed another way to understand the causal relationship of mind-body problem. Brains are the cause of minds in the sense of the H2O molecules are the cause of the liquidity of water. Obviously, H2O molecules and water are not two discreet events, but H2O molecules still are the cause of the liquidity of water. Therefore, the liquidity of water are caused by micro-particles and at the same time is the feature of these particles. That is to say, although the macro-level, the liquidity of water, is caused by micro-level, H2O molecules, these two levels do not appear one after another, they are two parallel levels. Searle regards this to be the right way to understand the relationship between brains and minds. Minds are caused by brains but they are parallel as different levels of description of one thing. "I am happy" and "My X neuron send Y electrical signals to Z neuron" are two descriptions on different levels of one same thing. However, unlike Spinoza's parallelism, which claims thoughts and extension are two different means to understand the same reality, Searle insists that there is causal relationship between minds and brains. Moreover, Searle claims the macro-level, minds, cannot be reduced to micro-level. In other words, people cannot say "X neuron feels hurt", just like saying "A H2O molecule is wet" is ridiculous. The problem is, Searle's own opinion of mind-body problem fires back at his argument against connectionism. By proposing The Chinese Gym and insisting individuals doing syntactical work in the gym cannot generate minds, Searle himself is saying that ridiculous proposition, brains cannot generate minds since a neuron cannot think independently. Therefore, if Searle agrees that minds as macro-level description cannot be reduce to neurons as micro-level description, he should also agree that even though individuals in the Chinese gym are unable to think independently, they could still generate minds as a whole. Or to express this point in a straightforward way, water is wet even though a H2O molecule is not wet.

In summary, by demonstrating two arguments against CRA, one from the Churchlands and the other one from Searle himself, the article try to draw the conclusion that CRA cannot confute connectionism AI models.

5. Conclusion

This review of The Chinese Room Argument elucidates the compelling nature of CRA and highlights its limitations when contending against connectionism. The conclusion of the compelling nature of CRA is that by linking semantics with intentionality, Searle made the premise (2) highly accords with people’s intuition of thoughts and minds. Specifically, he developed Austin’s theory of speech acts by replacing different types of speech acts to intentionality. He successfully connected semantics, a rather strange terminology, to something everyone has an intuitive feeling about. Therefore, Searle could use premise (2) to refute The System Reply and The Robot reply. However, CRA are weakened by The Luminous Room Argument and Searle’s own view of mind-body problem when it faces connectionism. The Luminous Room Argument invented by the Churchlands aims at Searle’s premise (2). By proposing a parallel thought experiment, the Churchlands thinks that “syntax is not sufficient for semantics” is not necessarily true since it is an empirical proposition instead of analytic truth. Moreover, Searle’s own view of mind-body problem thinks that although brains cause minds, minds and brains are still two irreplaceable levels to describe one thing. Therefore, just like water is wet although a H2O molecule is not wet, a connectionism network could have the possibility to generate minds although a unit which only does syntactical work cannot think or understand. There are numerous articles argue against, support or review the CRA, but few of them really think about the reason for CRA to cause our confusion or directly targeted at the core of CRA, the premise (2). This article reviews CRA from this special perspective. However, the article may lacks of more technical and current discussions about artificial intelligence, cognitive science and neuroscience. The article mainly focuses on philosophical issues caused by CRA. The rethinking of CRA and these philosophical issues caused by it can provide an inspiring prospective about the relationship between human intelligence and nowadays AI.

References

[1] *Searle, J. R. (1984). Minds, brains and science. Harvard university press.*

[2] *Cole, David, "The Chinese Room Argument", The Stanford Encyclopedia of Philosophy (Summer 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/sum2023/entries/chinese-room/>*

[3] *Hornsby, David (2014) Linguistics: A Complete Introduction. Teach yourself books. Hodder & Stoughton, London, UK. 134.*

[4] *Jacob, Pierre, "Intentionality", The Stanford Encyclopedia of Philosophy (Spring 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/spr2023/entries/intentionality/>*

[5] *L. Austin. (1975) How to Do Things with Words. Oxford: Oxford University Press, 94-103.*

[6] *Searle, J. R. (1983). Intentionality: An essay in the philosophy of mind. Cambridge university press.*

[7] *Dinsmore, J. (2014) Thunder in the Gap. In: Dinsmore, J. The symbolic and connectionist paradigms: closing the gap. Psychology Press. 1-2. 2014.*

[8] *Buckner, Cameron and James Garson, "Connectionism", The Stanford Encyclopedia of Philosophy (Fall 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2019/entries/connectionism/>.*

[9] *Churchland, P. M., & Churchland, P. S. (1990). Could a Machine Think? Scientific American, 262(1), 32–39. http://www.jstor.org/stable/24996642*

[10] *Lee, Z. (2011). Viewing the Development of Artificial Intelligence from a Philosophical Perspective, A Critical inspection of The Chinese Room Argument. Journal of Henan Normal University, 06, 14-18. DOI:10.16366/j.cnki.1000-2359.2011.06.029.*