Can computers think? In attempt to make sense of this question Alan Turing, John Searle and Daniel Dennett put fourth varying arguments in the discussion surrounding the possibility of artificial intelligence. While this inquiry at first appears to be solely one of science rather than a feature of philosophy, the ideas of the Turing Test, the Chinese Room, and True Believers have proved themselves influential to the study of the mind. I aim to critically discuss these ideas as well as explain Dennett's position on the arguments of Turing and Searle.

The Turing Test is Alan Turing's attempt to clarify the conversation. For Turing, the question of the possibility of mechanical thought is—at least as of 1950—"too meaningless to deserve discussion" (Kim). Turing argues that a more profitable route of inquiry is to instead devise a methodology from which we can attribute computers "the kind of mentality that we grant to humans" (Kim). The best methodology, Turing claims, is to play the Imitation Game: an interrogator communicates blindly with both a computer and a human and is asked to identify which responder is the computer. The ability of a computer to perform in the Imitation Game is its propensity to fool the interrogator into thinking it is human. A computer is then said to pass the Turing Test if it consistently fools the interrogator in the Imitation Game and vice versa. Thus, by passing the Turing Test, the computer is said to be intelligent. Alan Turing believed that the question of machine intelligence would, by our current time, be passé, such that "one [would] be able to speak of machines thinking without expecting to be contradicted" (Kim).

Before mentioning any criticisms of the Turing Test, I would like to distinguish that I interpret Turing as *not* viewing his test as a universally sufficient condition for computer intelligence but instead as a rule-of-thumb and, more importantly, a practical process for establishing a criterion of machine intelligence. It is obvious from the description of the Test that it is in some part inherently subjective; the Test relies on the use of a specific individual to play

interrogator, giving way to a biased performance in the Imitation Game. Furthermore, the Turing Test is itself somewhat ambiguous in order to shift focus away from the specifics, such as the length of time used for playing the Imitation Game and the medium of communication (for they are dependent on the constant flux of technological innovation), and towards the general idea of formulating a method of determining a computer's intelligence. I believe that this important because it should also shift criticisms of the Turing Test away from the specifics and towards the general.

My criticism against the Turing Test is that the general idea of the Test seems to be one of determining a computer's *imitation* of intelligence rather than intelligence itself. If it is the case that the Turing Test examines the ability of a computer to imitate a human rather than to be a human, then we must also question whether pure imitation of intelligent behavior is really identical to intelligence. Is all that there is to having a mind really just acting as if you have a mind? This is the distinction raised by Searle with his terms Strong AI and Weak AI and it applies directly to the Turing Test. Put succinctly, my complaint with the Turing Test is that it does not tell us if a machine has a mind, but only if a machine acts as if it has a mind (i.e. if it has Weak AI). Alan Turing's view presupposes the idea of machine functionalism, or that our mentality and intelligence are the result of the brain's operation as a computer. In other words, Turing believes that the only thing to having a mind is running the right computer program using the right inputs and outputs. It is through this idea that the connection between a machine acting as if it has a mind and a machine actually having a mind is made.

Criticisms of the Turing Test lead us to another question: if we reject the Turing Test as a method of determining computer intelligence, then what method *can* we use as sufficient for claiming intelligence? In our day-to-day lives, we apply a sort of Turing Test to conclude that

squirrel that we see on our way to class is indeed not intelligent in the same sense as the professor who is going to give the lecture. How do we come to this conclusion? For most of us, it seems, we interact with the two entities and draw conclusions from their ability to respond to us. We use the very same sort of analysis as the Turing Test, namely communication and response, to judge if the entity has a mind. In both the everyday scenario and the Turing Test, we make the connection between X acting like it has a mind and X actually having a mind. While this method works in practice, the justification for this connection between imitation and mind isn't sufficiently established by Turing through machine functionalism. Turing does not give adequate justification for why we should accept machine functionalism, which has turned out to be pivotal in his Turing Test.

John Searle addresses the issue of creating a sufficient condition for establishing artificial intelligence by arguing that the very existence of computer minds is impossible. Searle defines Turing's functionalist view, "that the mind is to the brain, as the program is to the computer hardware," as Strong AI (Searle). Searle characterizes this view as rejecting anything essentially biological about minds such that all there is to having thoughts and feelings is implementing the right program on the right hardware. Searle's syntax argument against the existence of Strong AI goes as follows:

(1) By definition, computers have purely syntactical (i.e. formal) operations
(2) By definition, minds and mental states require semantic operations
(3) Syntactical operations are not sufficient for semantic operations
(4) Computers cannot have semantic operations
(5) Therefore computers cannot have minds and mental states

When Searle talks about the mind, he means to assume that it must have semantic content. Searle states that, *by definition*, "there is more to having a mind than having formal or syntactical processes" (Searle). If X is without any mental content *about* something, or without

any mental content with meaning (as opposed to purely symbolic manipulation), Searle asserts that X does not have a mind. While I'm not going to discuss this point in detail, I think that this definition of the mind could possibly be challenged by Alan Turing and other machine functionalists. Additionally, Searle views this argument as independent from any state of technology because its nature "has to do with the very definition of a digital computer, and what a digital computer is" (Searle). As long as we continue to define computers by their operation via syntactical, symbolic manipulation, Searle believes that his argument will still hold.

One response to Searle's syntax argument is to reject that (3) syntactical operations are unable to produce any semantic properties. While a computer may only have syntax for its internal operation, it seems possible that the very same computer could also produce some sort of semantics by being interpreted by external entities. By the syntax being interpreted as semantics by the external entities, the entire system gains semantic properties. And sense the entire system has semantic properties it could also possibly have a mind—including the purely symbolic computer. The most apparent response that Searle might give is that such a system in itself (i.e. internally) does not have semantics; the only semantics are those originating from the interpreter. The system itself is still internally purely symbolic. The system objection does not seem to be much of a threat to Searle's argument and it is explicitly addressed within his Chinese Room thought experiment and argument.

A question that I would like to raise to Searle before discussing his Chinese Room argument is a concern regarding the real benefit of rejecting Strong AI. Even if we completely accept Searle's point of view, it seems very possible that computers are capable of acting exactly, or near-exactly, as if they did have a mind! Is this not artificial intelligence in the pragmatic sense of the term? Technology isn't looking to create robots that *are human* in the mental sense,

but rather robots that *are capable of acting as if they were human*. Searle himself concedes the possibility that a robot could act as if it had a mind without necessarily having Strong AI. Why should we care that computers are incapable of having minds if they *are* capable of acting indistinguishably as if they did? It seems that, if this is the case, we need not worry about the impossibility of Strong AI. All we need to do is change our language a bit: instead of saying that the thermostat has beliefs, we say that the thermostat acts as if it has beliefs. The issue then becomes whether a computer can really act as if it has a mind even if it doesn't really have a mind, to which I would expect Searle to contend that it cannot.

To demonstrate his argument, Searle presents his famous Chinese Room thought experiment. Suppose the existence of a computer program that simulates an understanding of Chinese such that, given a question in Chinese, the computer will match the question against its data base and provide an answer as good as a native Chinese speaker. Now imagine that you are locked in a room with the same syntactical program instructions in a language you can understand. You can now communicate as if you understood Chinese to the external world, but do you really understand Chinese? Searle states that you do not. Instead, you are limited to simply reading and executing the complex set of instructions given to you with no actual understanding of what is being communicated. The important contrast between understanding versus executing (or imitating versus duplicating) is manifest if we compare the scenario to communicating within a language we do understand, such as English. Now "you understand the questions in English because they are expressed in symbols whose meanings are known to you" whereas in the Chinese case you had simply manipulated symbols according to a set of formal instructions, without ever attaching any meaning (Searle).

It now makes more sense to talk about the systems criticism. Searle states that within the system, there still must be a central processing unit of some sort that acts exactly as described by the Chinese Room. In other words, there is no way for the entire system, even if taken as a totality, to have semantics because it would still require the central processing unit, or the Chinese Room, to get from syntax to semantics. Searle provides another example: suppose that you have memorized the entire set of complex instructions found in the Chinese communication program. You are now an entire system, but you still do not understand Chinese because you still only know the set of instructions and nothing about the actual semantics of the characters that you are transcribing. While I think that Searle's response to the systems criticism is sufficient, one question that could arise is what then does constitute a meaningful understanding of a language. Why is Searle so certain that our understanding of English is itself not solely dependent on the memorization of a long, complex set of rules?

Another response against the Chinese Room argument is the Robot criticism. Now imagine that we have "put the Chinese understanding program inside a robot. If the robot moved around and interacted causally with the world, wouldn't that be enough to guarantee that it understood Chinese" (Searle)? Once again, Searle shows us that, as long as we suppose that the robot has a computer for a brain, it would still run into the issue of the Chinese Room, or the inability to get semantics from syntax. This criticism can also be combined with the entire systems criticism; what if the robot interacted casually with the world and was considered as an entire system? Wouldn't the robot then, as a whole and including its central processing unit, understand Chinese? I think that Searle's previous systems response to this question, namely the reference that *somewhere* we are limited to solely syntactical operations, still applies to this combined criticism. But what I do think this combined criticism brings to light is the very

question I posed to Searle previously: if we really do have a robot running around and interacting casually with the world as if, for all intents and purposes, it did have a mind (or understood Chinese), then does it really matter that it doesn't have a mind (or doesn't understand Chinese) in the *strict* (Strong AI) sense? Sure, the robot doesn't actually think, but it acts as if it does! This type of issue is not discussed by Searle within the Chinese Room argument.

Daniel Dennett takes a different approach to discussing artificial intelligence, beginning with a description on what it is to have beliefs. While we attribute beliefs metaphorically to diverse things, Dennett claims that "what it is to be a true believer is to be an intentional system" (Dennett). Dennett goes on to explain that an intentional system is that which can be reliably predicted using the intentional strategy. But what is the intentional strategy and how do we use it? The intentional strategy is, stated simply, to treat "the object whose behavior you want to predict as a rational agent with beliefs and desires and other mental states exhibiting… intentionality" (Dennett). To apply the intentional strategy, we first (1) assume that X is a rational agent (which can be further detailed given additional empirical data), then (2) attribute reasonable beliefs and desires to X that it ought to have based on its sensory experience with its environment and a basic set of needs ("survival, absence of pain, food, comfort, procreation, entertainment"), and finally (3) predict the behavior of X assuming that it will work towards what it desires based off its beliefs. Dennett asserts that the intentional strategy successfully models the behavior of a wide variety of agents, including the simple behavior of lightning and thermostats to the complex behavior of human beings.

Clearly the intentional strategy works well at approximating the behavior of a diverse set of objects, but Dennett argues that "even when we are surest that the strategy works *for the wrong reasons*, it is nevertheless true that it does work, at least a little bit" (Dennett). How is

Dennett sure that his system always works? For simple objects like thermostats and lecterns the answer is that they are well designed. For more complex objects like human beings the answer is a bit more complicated. While we might have theories for why humans behave the way they do, the reality is that Dennett claims we just don't know why the intentional strategy works. Dennett wants us to accept his intentional strategy due to the empirical evidence that the system has worked in the past without providing much explanation of why the system worked in the first place. Furthermore, Dennett wants us to understand his idea of true believers as one of degree. We should avoid trying to divide objects into believers and non-believers and instead understand that the real difference is the degree of demand for each object's representation of the world.

It doesn't take much critical interpretation to see that Dennett sides with Turing over Searle. I think that Dennett would regard the Turing Test as a very reasonable way of attributing a sort of intelligence to objects. Anything that passes the Turing Test would also be reasonably explained as an intentional system. Dennett explicitly describes a chess-playing computer as a viable candidate for his intentional strategy, indicating to me that surely any computer that can convince an interrogator that it is a human being is also a practical candidate. The issue I have with both Turing and Dennett is that while their methods are great for determining if a computer acts *as if* it has beliefs, or mental states, or a mind, I am not as convinced that this is proof that the computer has a mind in the identical sense as an adult human being. It seems to me that the non-human intentional systems are simply lacking some of the key characteristics of human minds, such as consciousness and mental content about something.

On the other hand, Dennett would not regard Searle's Chinese Room argument as favorably as the Turing Test. Dennett might argue that if a computer, system, or robot is able to act as if it has a human mind, then it has just as much of a right to a human mind as any human.

If we were to apply the intentional strategy to the computer, system, or robot and compare our results to the application of the intentional strategy of a human being, I think Dennett would predict them to be nearly identical. For Dennett, this is all that is needed to end the discussion on artificial intelligence: if we consider ourselves intelligent, then the computer, system, or robot that is classified as a similar intentional system to ourselves must also be intelligent. The issue of syntax and semantics goes away because Dennett isn't worried about whether the computer, system, or robot *actually* has beliefs and desires, but only if the computer, system, or robot has behavior that can be predicted *as if it had* beliefs and desires. While I disagree that all there is to having a mind is a function or an imitation, I think that Dennett's perspective is a much more pragmatic way of talking about artificial intelligence. The scientist need not worry that his robot cannot actually think; only that it can imitate thought so that it can fulfill its purpose. With the right hardware, wetware, or meat it seems that artificial intelligence, in the weak sense, is absolutely possible within the views of both Searle and Dennett.