Dan Ports 2002/12/10
21L.448 (**revised**)

A Rebuttal to Searle's Chinese Room Argument

The ultimate goal of artificial intelligence research is to create computer systems that can simulate thought at a human level. The task of creating such a system encompasses a variety of technical challenges. It also prompts a philosophical question: if such a system were created, how much intelligence can be attributed to it? Some have claimed that it must be capable of "thinking" or "understanding" in the same manner as a human, while others claim it is simply blindly following rules. John Searle, in his paper "Minds, Brains, and Programs," argues against the notion that "the appropriately programmed computer really *is* a mind" (353). He uses his famous Chinese Room example to claim that "intentionality" is the difference between a thinking human and a computer that is merely processing information according to rules. According to Searle, though the machine may be behaviorally indistinguishable from a human, it is impossible for anything that does not have "the same causal powers as brains" to be intentional, and therefore it cannot be intelligent (369). This argument is flawed because the claim that only human brains are capable of intention is not as trivial as Searle makes it appear. When considered from the perspective that humans are a form of biological machine, Searle's notion of intentionalism becomes an arbitrary, unclear distinction.

One definition for intelligence is based on observable behavior: a system is intelligent if it behaves intelligently. This is the definition proposed by Alan Turing in his paper "Computing Machinery and Intelligence." In this paper, he introduces the Turing test, which involves allowing a human judge to communicate via a text-based terminal with a computer program and another human. The judge is not told the identities of the two others, and the computer is programmed to imitate a human; the judge must determine which is really the human and which is the machine. If the two cannot be distinguished, then it follows that the computer must have mental abilities comparable to those of a human. The power of the test comes from its ability to test such a wide

variety of subjects. Turing's example transcript shows that a conversation can test knowledge of subjects as diverse as poetry, arithmetic, and chess; he claims that "the question and answer method seems to be suitable for introducing almost any one of the fields of human endeavour" (6). In order to pass the test, the computer must have a strong foundation of general knowledge and human-level ability in different fields; if it does not, the Turing test will expose it. According to Turing, a computer that can successfully pass the test must have "the intellectual capacities of a man," and it can reasonably be said to be intelligent and "thinking" (5).

John Searle accepts Turing's claim that machines can eventually be made capable of passing the Turing test. However, he rejects the notion that this will make them intelligent. He illustrates his argument using a famous thought experiment: a man who does not speak Chinese is placed in a room and given a book of rules in English which he can understand. He is then passed a set of Chinese symbols, and he follows the instructions in his book in order to assemble a response from Chinese characters. By doing this, he is responding coherently to questions posed in Chinese, and indeed it is possible that, with a large enough rulebook, he would be able to pass the Turing test in Chinese. But Searle claims that "I have inputs and outputs that are indistinguishable from those of the native Chinese speaker … but I still understand nothing" (356). By this reasoning, a computer cannot be intelligent: it is simply performing "computational operations on purely formally specified elements, [which] by themselves have no interesting connection with understanding" (357). Searle argues that a system is intelligent if it is capable not just of giving the correct output but also understanding the meaning, and this requires intentionality. This creates a dichotomy between conscious, intelligent systems with intentionality, and machines that are merely manipulating symbols. He rejects the "dualism" of strong AI, which postulates that mind and brain are different, and "programs are independent of their realization in machines" (371). For Searle, the mind and brain cannot be separated, because an intelligent mind must have intentionality and intentionality is purely "a biological phenomenon" exclusive to the brain (372). Thus, the only way a system can possibly be intelligent is if it has a nervous system like that of a human. Searle writes that, though they can be used to construct computing machines, "stones, toilet paper, wind and water pipes are the wrong kind of stuff to have intentionality in the first place — only something

that has the same causal powers as brains can have intentionality" (369). By this reasoning, a computer program, even if it meets Turing's behavioral standard for intelligence, cannot be intelligent by Searle's standard of intentionality.

A major problem with Searle's argument is that it draws an unreasonable distinction between the human brain and machines. This can be best understood by considering the perspective that the human body is a form of machine. This may seem to be a shocking idea, since so many philosophies and religions hold that humans, animals, and machines are in some respect fundamentally different. However, in a physical sense, they are more similar than they might seem. A machine is a collection of parts assembled into a functional system. This can also be said of the human body: it is composed of a set of organs that operate together, and these organs are in turn made up of cells, increasingly smaller parts. The major physical difference is that a human is a machine composed of biological parts, whereas a man-made machine is generally made from wood, metal, or similar materials. Yet when they are reduced even further to the atomic level, both humans and machines even share the same constituent elements; they are simply arranged differently. This applies as well to the brain. Searle claims that the human brain is intelligent because it is capable of understanding the input it manipulates. He acknowledges that a machine with an artificial nervous system would be equally intelligent, and that "it might be possible to produce consciousness, intentionality, and all the rest of it using some other sorts of chemical principles than those that human beings use" (368). Searle claims that the only way a computer can be intelligent is if it simulates a biological brain. This claim seems unreasonable because it implies that there is some intrinsic property unique to the brain that makes it capable of understanding. The brain is simply a machine made up of neurons in some particular complex arrangement, which could conceivably be replicated on a sufficiently powerful computer. Searle admits that such a system would be intelligent. But this arrangement is simply a system that takes inputs and generates an output by processing the data through neurons. The neurons are biological computational elements, and there is no reason that they could not be replaced with digital computational elements. If this is done, then the brain has been transformed into a program for a digital computer of the very sort that Searle claims cannot be intelligent. But the two are equivalent,

and it cannot reasonably be said that one is intelligent and the other is not. Searle's assertion that only a brain can be intelligent is flawed because the brain is simply a computing machine that happens to be implemented using biological components. The distinction he draws between biological and mechanical or electronic implementations is not nearly as significant as he claims it to be.

Searle's argument centers on the notion of intentionality: that the human brain can have intention, but computer programs cannot. Unfortunately, this notion of intentionality is not a well-defined, meaningful concept. Intentionality is not a natural, intrinsic property of an entity, but an arbitrarily assigned label that depends on perspective. Norbert Wiener suggests the impossibility of assigning intention when he writes in *God and Golem, Inc* that "a hen is merely an egg's way of making another egg" (36). This situation described in this statement is not an unusual one; the cycle of "biological alternation of generations" in which an egg produces a hen and a hen produces an egg is well understood. However, it is surprising to see it phrased this way because we expect it to be the hen rather than the egg that has consciousness and intention; we expect that the hen produces a egg in order to fulfill its goal of creating another hen, not vice versa. Yet either viewpoint can be valid. When viewed from an evolutionary perspective, both the hen and egg are equally necessary for the species to continue to exist, and each one creates the other; the two exist as a "duality" (35). We normally attribute intention only to the hen, because a hen can perform various actions and exhibit mental states. But Wiener is suggesting that the egg can also be seen as having intention, because, even though it does not directly act on the world, it is capable of producing a hen, which does. The choice of which has intention is therefore somewhat arbitrary; it is not a natural distinction so much as it is an artificial one that reflects the perspective of the observer.

Even Searle agrees that intentionality is a quality arbitrarily assigned by humans when he says that "we find it natural to make metaphorical attributions of intentionality to [our tools] … but no philosophical ice is cut by such examples" (358). Searle claims that the purpose and intentionality of tools is merely a metaphor we have ascribed to them for our convenience in describing them. We may view them as having intention, but this is meaningless; it is not a real

form of intention, as it certainly does not make the tools conscious or intelligent. If we accept this, then we must question whether all other intentionality is equally meaningless, including the intentionality we assume the human mind to have. If we again treat the human brain as a machine made up of biological components, we observe that the individual neurons do nothing more than transmit and combine signals; they cannot have individually have intentionality or consciousness. The human mind comprises an arrangement of these neurons. Applying the rules of physics, we could conceivably predict the thought or action the brain would produce in response to any particular stimulus: it is an output determined entirely by the reaction of the brain's neurons to a certain input.

A similar argument that human behavior is deterministic is found in the Book of the Machines in Samuel Butler's *Erewhon*. The Erewhonian author claims that free will is merely an illusion, using the example of a train driver who "can stop the engine at any moment that he pleases, but he can only please to do so at certain points which have been fixed by …an unseen choir of influences, which makes it impossible for him to act in any other way than one" (218). The Erewhonian uses this to argue that humans are similar to machines, because both behave deterministically not "spontaneously." In this regard the brain is essentially the same as a computer: it is a machine composed of elements that lack intentionality, and it produces a deterministic outcome according to a set of rules. Thus if we accept Searle's premise that a computer program lacks intentionality, we must say that the same is true for the human brain. This definition of intelligence based on intentionality suggests not only that are machines not capable of being intelligent, but that neither are humans. This result is absurd; it is not consistent with our intuitive definition that humans are intelligent, nor is it very helpful for assessing AI development. Therefore, intentionality is not a useful distinction for evaluating intelligence.

These problems with Searle's argument can be illustrated by comparing the Chinese Room example that he describes to a similar system in which the man in the room is actually capable of reading and speaking Chinese. We can suppose that the two are behaviorally identical, that each can give the correct answers to questions posed in Chinese. Searle would argue that the two are fundamentally different: the Chinese speaker is intelligent because he has intentionality and actually

understands the questions, while the other man lacks intentionality and is simply manipulating symbols. This argument sounds compelling, but upon further examination, the differences between the two are not as significant as Searle believes them to be. The first difference is an obvious one: one man requires a rulebook to produce the correct output, and the other does not. The difference is addressed by the "systems reply," which suggests that "understanding is not ascribed to the mere individual" but to "the whole system" consisting of both the man and the rulebook (358). Searle claims that it is "implausible" because if "a person doesn't understand Chinese, somehow the *conjunction* of that person and bits of paper might understand Chinese" (359). This reasoning is flawed: it is certainly possible for a system as a whole to have intelligence that its individual parts do not. After all, the individual neurons that make up the native Chinese speaker's brain are not intelligent by themselves, but the brain, the system that is their combination, is capable of understanding Chinese. Indeed, the human brain could also be reasonably described as a processor with an internal "rulebook" of sorts that is constantly updated the knowledge it has learned. A machine can similarly update its rulebook; this is the basis for the process of machine learning, which Norbert Wiener refers to as "a machine constructing another after its own image" (30). So the "systems reply" is a reasonable approach for assessing the situation; the man and rulebook must be considered together as a system that can have more intelligence than the man alone.

As another rebuttal to the systems reply, Searle proposes that "the individual internalize all of these elements of the system … memorizes the rules in the ledger and the data banks of Chinese symbols, and he does all the calculations in his head" (359). In this case, Searle is correct in claiming that the man is now the entire system, and he still does not understand Chinese. The man does, however, understand how to read, interpret, and reply to stories in Chinese, which is a large part of understanding the language. All that he is missing is the ability to associate the words with the concepts they represent; he, like a computer, "has a syntax but no semantics" (370). Searle believes that it is not possible for a computer to have semantic understanding, that a program's "first-order symbols don't have any interpretations as far as the computer is concerned. All the computer has is more symbols" (370). But the semantic meaning of words is simply another layer

of knowledge atop the formal symbols of language. To be able to understand Chinese, the man in the room needs to be able, for example, to connect a word like "hamburger" with knowledge about hamburgers, associated memories, and sensory input such as a visual image of a hamburger. This could simply be another set of rules and information added to the rulebook; there is nothing special about this type of knowledge that prevents it from being given to a computer. If the man in the Chinese Room is given a set of semantic knowledge in his rulebook, and he internalizes the rulebook, then he can be said to actually understand Chinese. The man-with-internalized-rules system will then not only be behaviorally indistinguishable from the native Chinese speaker, but they will also be mentally indistinguishable.

Searle's premise is that a system may behave intelligently — as measured by the Turing test, for example — without actually being intelligent. He claims that in order to be intelligent, it must be consciously capable of understanding what it is doing, which requires intentionality. This argument is flawed because the notion of intentionality is really an arbitrary distinction rather than a natural property, as Wiener's example of the egg and the hen helps demonstrate. He further argues that intentionality is an intrinsic property of the biological brain that makes it capable of intelligence, and that a computer system, not having the same implementation as the brain, cannot be intelligent. This argument is refuted by considering the brain as a biological machine. When it is treated simply as an organized collection of neurons, the brain is a deterministic information processing system that maps sensory inputs to output actions, very much like a computer program. This implies that a system can still be intelligent if it is implemented electronically rather than biologically. Of course, this does not mean that achieving artificial intelligence is a simple goal. Alan Turing's prediction of the development of machines capable of passing the Turing test within fifty years has not proven true. Much research is still ongoing to determine what the appropriate rules for such a system should be and what the best structures for representing knowledge are. However, the obstacles that currently prevent us from having thinking machines are technical ones, not the more fundamental philosophical objections Searle proposes.