

Connectionism and the Chinese-Room

About twenty years ago, John Searle began a huge body of discourse with his rejection of a computational theory of mind. Searle is arguing against "strong AI", his term for computational theory of mind. Strong AI would claim that an appropriately programmed computer can possess cognitive states, that it really is or has a mind.

Searle is arguing generally against a computational approach to mind. To this end, he utilizes a now infamous thought experiment, known as the Chinese-Room. At the core of Searle's thought experiment there is a human simulating a computer. The human in the room is given meaningless symbols in Chinese script and instructions in English for manipulating them according to their shapes. By following the instructions and correlating the symbols he appears to those outside the room to understand the Chinese language. In reality he understands nothing; he is just manipulating the symbols according to their syntactic properties. Since he is doing in essence what a computer does and he doesn't understand Chinese, then no computer can ever come to genuinely understand Chinese, or anything for that matter. At least, not the way humans do.

To understand the effects of Searle's argument, we must first understand what exactly he is arguing against. Searle directs his efforts at what he calls Strong AI. At the time, this term encompassed all computational theories of mind. To fully analyse the effects of Searle's experiment, it is helpful to distinguish two separate types of computational theory. This will be done by examining the use and dismissal of standard symbolic atoms.

The first is known as the Classical Cognitivist view. This is arguably the most popular version of computational theory, and has undoubtedly produced the most literature. Key to this view, as Searle has shown in his thought experiment, is the manipulation of symbols. Classical

systems maintain that the fundamental aspects of intelligence are found in symbol manipulation. This view rests on the assumptions Newell and Simon state collectively as the “Physical Symbol System Hypothesis.” This states, simply: a physical symbol system is any system in which suitably manipulable tokens can be assigned arbitrary meanings and by means of careful programming be relied on to behave consistently (Clark 1989). To Newell and Simon, the most important property of a symbol is that it represents something. Further, symbols are considered to be atomic, that is, they are indivisible. Symbols can be combined together to form expressions, but they cannot be broken down. A symbol therefore is an indivisible unit, designating something, which can be manipulated by a physical symbol system, and lead to intelligent behaviour. These symbols are vital to the operation of classical systems as they are the fundamental units of computation. .This symbol system has formed the basis of most of the computational research into AI. The paradigm formed around it seems vulnerable to Searle’s line of attack.

If Classical systems can be characterized as operating in the symbolic paradigm, or computing with symbols, than Connectionist systems can be classified as operating in the sub-symbolic paradigm or computing *on* symbols. In a Classical system the level of computation system is found on the same level of the symbols, or atoms. Alternatively, the level of computation in a connectionist system occurs below the level of the symbol. In the sub-symbolic paradigm, cognition is not modeled by the manipulation of machine code that neatly matches our symbolic descriptions (words). These descriptions are only labels that bear approximate relations to the underlying computational structure (Clark 1989). Connectionist systems could be considered analogous to how we humans understand words like ‘kitchen’ or ‘chair’. Although these words correlate to actual objects in the real world, our understanding is made of a vast

array of details. For example 'kitchen' is a single symbol, but underlying it is a vast network of conditions we need satisfied for something to *be* assigned the symbol 'kitchen' and not something else, like 'bedroom' or 'restroom'. This system of underlying conditions is the computational level of connectionist systems. The essence of connectionism is a rejection of the use of symbols, at least in the sense that a classical system would use them. Rather than having symbols that map from the computational level directly onto the representational level, the representations have vast numbers of activation patterns on the computational level that correlate. It is in this fundamental difference from classical systems that a computational theory of mind may avoid the brunt of Searle's Chinese room argument.

The way Searle's argument can be used against classical computationalism is clear. Computer programs may manipulate symbols but they will not have any understanding of what those symbols correlate to outside the program. Searle's human, inside the room is proof of this lack of understanding. Searle asks us to compare the how the human sees the Chinese symbols, and the instructions for manipulating them. This is Searle's proof against computationalism. Namely that human-like understanding cannot be derived from symbol manipulation, no matter how complex.

This thought-experiment seems to have classical systems in a bind, the fundamental symbol manipulation these systems rely on is incapable of recreating human understanding. And therefore is a generally useless attempt at modeling human cognition. How does this line of argument stand up against connectionism?

Connectionist unlike, classical models have a different way of viewing symbols. In the classical system, units of computation *are* the units of representation, in connectionist systems,

the units of computation and representation are both present, but they are not the same. This distinction avoids most of Searle's argument.

First, consider the units of computation. It is true that these are featureless entities in connectionist systems, with nothing to them except perhaps a unique shape or degree of activation. Using Searle's argument, these units cannot lead to human like understanding. But these units were never meant to produce human understanding. At least not directly anyways. Unlike their counterparts in classical systems, these units are only intended to be syntactic objects. As is made clear by the symbolic/subsymbolic distinction, the computational level in connectionist systems is not a representational level. Therefore Searle's argument, applied to the units of computation, tells us nothing of importance.

Second, let us consider representations. Representations in connectionist systems, are not manipulated directly, but rather are the indirect result of low-level manipulations. Most importantly, connectionist representations have a rich *internal structure*. Unlike symbolic representations, connectionist representations have their own intrinsic organization, by virtue of their consisting in a complex pattern of activation. These representations are totally unlike those of classical systems. It would seem that Searle's argument cannot apply to connectionist systems in a fundamental way and therefore goes nowhere towards disproving it.

Searle would doubtless be unimpressed by the argument above. He might point out: (1) connectionist systems are still on some level computational, consisting in syntactic manipulation of symbols; (2) syntax is never sufficient for semantics; therefore (3) connectionist models can possess no true semantics. Even though the representations come from varied patterns and are not computed on directly, they result entirely from syntactic manipulations, and so cannot be said to ever model human like understanding.

A reply to this seems clear. The only proof Searle has in the semantics from syntax argument is found in the Chinese Room experiment. Because the thought experiment cannot apply to connectionist systems, neither can his proof. This does not show that connectionist systems can create human like understanding, simply that Searle's argument cannot be used against them.

This paper has examined the effectiveness of Searle's argument against what he calls Strong AI (computational theory of mind). The argument plays on a fundamental weakness in classical models: the fact that the units of computation are purely syntactic and that this cannot generate human understanding. Instead of treating this as a destructive argument, it is perhaps better to view it as a constructive criticism which AI might be able to overcome. AI research, chiefly through the connectionist paradigm, may be well on the way to dealing with this problem.