

**“Debugging Artificial Intelligence”**

**I. Introduction**

It could be argued that the ultimate goal of artificial intelligence research is to endow a synthetic being with human-level cognition. This dream has been the subject of both science fiction and academic inquiry for decades, and it was believed to be within our grasp ever since the digital computer was invented and introduced into society. However, given the current state of the field, it seems fair to claim that AI has failed to meet the expectations of the general public and the scientific community. Modern computers are very limited in what they can do and researchers have instead decided to focus on specific aspects of intelligence – like learning and perception – while segregating themselves into rival camps with conflicting methodologies.

One might wonder exactly why we have been unable to engineer a genuinely, or at least believably, cognitive artifact. Certainly not due to a lack of effort or innovation, as we have experimented with symbolic programming, neural networks, dynamical systems, artificial life, and many other avenues that we hoped would lead to intelligence. Perhaps in order to answer this question we must cast aside technological concerns and carefully examine the underlying assumptions behind human and artificial intelligence.

In this paper, I intend to explore the philosophical foundations of AI in order to explain the present status of the field and assess its future prospects. First, I will consider the classical model of AI – generally known as “computationalism” – and I will demonstrate its limitations by discussing John Searle’s contribution to the discipline, which is exemplified by his famous Chinese Room argument. Then, I will turn to a very different conception of the mental, one that avoids making any claims regarding the structure of intelligence and instead treats it as the exhibition of a certain kind of behaviour. I argue against this second approach as well, with reference to the work of Daniel Dennett and a brief mention of the Turing test and Rodney

Brooks' research. Finally, I will put forward a new philosophical framework that emerges from a constructive analysis of the previous two, where consciousness plays a central role in the kind of intelligence that AI once set out to reproduce.

## II. Searle's Legacy

*“The interest of the original claim made on behalf of artificial intelligence is that it was a precise, well defined thesis: mental processes are computational processes over formally defined elements. I have been concerned to challenge that thesis.”* (Searle 1980)

John Searle has clearly been a prominent figure in the philosophy of artificial intelligence. Most people would attribute his fame to the Chinese Room argument and his opposition to the AI doctrine of his time, but I think there is more to be said about the *nature* of his criticism. It is not a challenge to the technological practices of symbolic AI, but rather a philosophical assault on the *basic principles* of computational functionalism. To see this, consider the remarks often made by proponents of Connectionism. They might point out that digital computers have difficulty learning and that they perform poorly when faced with novel situations, while neural networks have no such problems. But this is only a concern at the level of implementation and could be remedied (at least to a certain extent) by the use of probabilistic algorithms, such as Bayesian learning, and other mechanisms. Most importantly, though, comments like these have nothing to say about whether computation *in principle* is the right approach to engineering intelligence.

Searle's critique, however, targets the very foundations of early AI research that was conducted since the 1950's. To the credit of pioneers like Newell and Simon, theirs was a clear hypothesis that rested on firm philosophical ground and had the right technological backing at the time – namely, the creation of the Turing machine. They also enjoyed a certain amount of success, so it would be foolish to dismiss their approach entirely. But when Searle came along, he formalized all the implicit assumptions of the field and coined the term “Strong AI” – referring to the computationalist position – which he articulates in the following manner: “there

is a level of mental operations that consists in computational processes over formal elements which constitute the essence of the mental, and can be realized in all sorts of different brain processes in the same way that any computer program can be realized in different computer hardware” (Searle 1980). What is important to note here is that thought is treated as computation, from which it follows that (1) intelligence is characterized by abstract reasoning and problem-solving, and hence (2) intelligence is a tangible property, possessed by some entities and not by others.

Having defined the rival view and identified its main tenets, Searle offers an argument as to why computationalism is essentially wrong about the nature of cognition. There seem to be two main premises that are worth examining in more detail. The first appears in “Minds, Brains, and Programs” (1980), where Searle introduces his famous Chinese Room experiment, and it is that formal symbol manipulation on its own is not sufficient for intentionality. More precisely, the conclusion is that the man in the room does not understand Chinese simply by being an instantiation of a computer program. If one were to combine this claim with a second premise stating that genuine cognition requires intentionality and “understanding” of language and other things, then one could make a compelling case against Strong AI by showing that *in principle* a computer could never *be* a mind – which is exactly what Searle does. Although I believe that this argument is generally sound, there appear to be some concerns regarding the first premise that cannot be ignored.

The conclusion of the Chinese Room experiment has been highly controversial – mainly because of its consequences, but also because of certain ambiguities in Searle’s line of reasoning. As Jack Copeland points out in “The Chinese Room from a Logical Point of View” (2002), the argument is not logically valid. Namely, one cannot declare that since the man in the room does not understand Chinese, the Room as a whole does not understand Chinese either. This sort of remark is similar to the Systems Reply, but it makes no strong claims about whether the Room does or does not understand Chinese – rather, it simply draws attention to the lack of logical

entailment from one proposition to another. As Copeland puts it: “One might as well claim that the statement ‘The organization of which Clerk is a part has no taxable assets in Japan’ follows logically from the statement ‘Clerk has no taxable assets in Japan’” (Copeland 2002).

Moreover, I think it might be worthwhile to assess the adequacy of Searle’s response to the Systems Reply. The essence of this objection seems to be that the man is one PART of a larger WHOLE, and should be treated that way – so we should not look to the PART to understand Chinese, but to the WHOLE. Searle suggests that the man should “internalize” all the other elements of the system, thereby reducing the WHOLE to the PART again, which preserves the argument, since it has been established that the man does not understand Chinese. However, it is not clear whether one part can be made to contain other parts of a system as trivially as Searle would like. For example, the engine of a car cannot consume the wheels, seats, etc., and then “become” the whole car. Nor can a computer’s Central Processing Unit “internalize” the hard drive and the rest of the components and then still be called a computer. So it remains to be determined if this response is acceptable given the concept of the Systems Reply. This is, of course, far from a definitive refutation of Searle’s conclusion – which I believe is correct – but it does raise some interesting points of debate.

Let us now return to the second premise of the original argument against Strong AI. The idea is that “meaning” and “understanding” is indispensable to human thought. Although there seems to be no dispute over this claim – and the general importance of intentionality to cognition for that matter – there is ample speculation regarding the nature and origins of meaning. Searle adopts the following view on the subject: “...only something that has the same causal powers as brains can have intentionality...” (Searle 1980). In fact, after he presents his case that computation is *not* thinking, he expresses strong doubts that formal symbol manipulation has *any* role in cognition. Instead, he says that we must study the neurobiological properties of the brain in order to figure out how brains *cause* minds.

Perhaps Searle dismisses the notion of computation prematurely, however. To see why, consider his intuitions behind the Chinese Room argument one more time. The obvious problem is that the man (or system as a whole) does not understand a word of Chinese. Searle also emphasizes that there is an absence of meaning (intentionality) in the symbol manipulation process. However, this statement is somewhat misleading. The squiggles on the pages *do* mean something to a Chinese speaker, just like the sentences in English mean something to the man in the room (after all, English letters are symbolic squiggles as well). So a more precise statement would be that the Chinese symbols don't mean anything *to the man*, in the sense that meaning is not derived from them by the agent (man or system). This seems to be the main problem and it is essentially equivalent to saying that the man does not understand Chinese.

In fact, Searle claims something similar – that in the Chinese case the man produces answers by shuffling around *uninterpreted* formal symbols (Searle 1980). He also points out that the intentionality computers appear to possess is entirely derivative – that is, computers require a user to interpret their functioning – which leads him to reject computation as having anything to do with thought. However, he does not seem to consider the possibility that formal symbol manipulation *does* play a role in “thinking” by forming the basis of human-level cognition, and what is absent is a derivation of meaning by the agent in question. This is not an implausible hypothesis, especially since it seems to follow naturally from our discussion of the Chinese Room and the problems of “meaning” that it raises. I will develop this idea further when I attempt to provide a new philosophical framework for AI, but now let us turn to the notion of “derivative intentionality” and another view of intelligence that is very different from that of computationalism.

### III. Dennett and Derivative Intentionality

“...*the intentionality of our mental states and processes is derived in just the same way as that of our books and maps (and the inner states of our robots).*” (Dennett & Haugeland 1987)

In the previous section I examined a theory of mind that treats thought as computation over formally defined elements in a physical symbol system. This outlook was adopted by most AI researchers in the early stages of the field’s development and it remains popular to this day. Perhaps its appeal is best explained by the fact that it makes specific claims regarding the nature of intelligence – that it is an *intrinsic* property of certain entities and consists of problem-solving using abstract representations of the world. As Searle has shown in his work, the intentionality possessed by such machines is strictly derivative, which means that it only exists in the minds of external agents who interpret the behaviour of computers in some way. This is clearly a problem according to Searle, on the basis of which he concludes that the computationalist position is faulty. However, not all philosophers of AI agree with Searle’s assessment of the situation. There is another approach to explaining mental phenomena that only focuses on the behaviour of the entity in question. Proponents of this view believe that all intentionality and intelligence is derivative, which is why they would disregard Searle’s objections to symbolic AI. I will now briefly consider the work of Alan Turing and Rodney Brooks in order to illustrate the main tenets of this behaviour-based framework and then I will discuss its adequacy by looking at the views of Daniel Dennett and Searle’s opposition to them.

It seems rather peculiar that Turing would subscribe to a doctrine other than computationalism, since he essentially laid the foundation for symbolic AI research. However, his famous “Turing test” reveals his preoccupation with the behaviour of the system, rather than its internal processes. What really matters is whether the machine can perform as well as a man in the “imitation game” (Turing 1950). As Searle puts it: “The Turing test is typical of the tradition in being unashamedly behavioristic and operationalistic” (Searle 1980). The real merit of this test, then, is that it allowed scientists to even entertain the possibility of non-human

intelligence, which opened the door for AI research in the 1950's. In fact, the imitation game is so general that one could replace the man with a thinking cactus, instead of a Turing machine, and still get the same outcome if that cactus exhibits intelligent behaviour, as perceived by a human interrogator.

The robotics research of Rodney Brooks is rooted in the same philosophical principles as exemplified by the Turing test. He argues against the notion of central representation and insists that intelligence emerges from the interaction of the system's components. Hence, he is not so much concerned with the structure of thought, but with engineering a creature that exhibits the right behaviour – specifically, one that would be *interpreted* as intelligent. After all, intelligence is in the eye of the observer (Brooks 1991).

Daniel Dennett is also a major advocate of the behaviour-based approach to cognition. For example, he seems to remain agnostic regarding the actual existence of propositional attitudes (beliefs, desires, etc.) in the mind and instead claims that “rationality” consists of behaving in a way that is consistent with what he calls the “intentional stance” (Dennett 1981). Therefore, if an observer was to adopt this explanatory position and treat an object *as if* it was a rational agent, then the successful prediction of that entity's behaviour would indicate that it is indeed a “true believer” – that is, a thinking thing – regardless of the object's internal composition. Moreover, Dennett maintains the same attitude towards intentionality as well. He argues that any mental representations that exist in our heads have the same status as trails of ink on paper – their meaning must be derived. And so he claims that there is “...no more real, or intrinsic, or original intentionality than that” (Dennett & Haugeland 1987). In summary – it's all derivative for Dennett and for many others who embrace this philosophical framework.

The question, then, is what we should make of all this behaviour-talk. Is intelligence and intentionality only real insofar as an observer makes it real? It seems that this view is nearly correct, in the sense that meaning certainly *is* derived by a human agent, as is the case with computers and their rich (but derivative) semantics. Searle agrees with this point. But where

Dennett et al. are mistaken in that there is no such thing as original intentionality and that one can only judge the intelligence of a given entity by its behaviour. Searle comments on this claim: “And the mental/nonmental distinction cannot be just in the eye of the beholder – it must be intrinsic to the systems. For otherwise it would be up to any beholder to treat people as nonmental and, for instance, hurricanes as mental, if he likes” (Searle 1980). Moreover, Dennett’s position may be somewhat self-refuting – that is, even if intelligence is exhibited by the actions of an object, there still needs to be an agent that is sophisticated and cognizant enough to be able to determine what counts as intelligent behaviour. So it also seems illogical to claim that all intentionality is derivative, since the term “derivative” implies the existence of a “derivator”, which just leads to the question of how this “derivator” is able to extract meaning from objects in the first place. In other words, without some sort of explanation of the origins of intentionality, there will always be an infinite regress of derivation.

In the next section, I will attempt to arrive at a synthesis of the major philosophical frameworks examined in this paper. More specifically, I will consider the role of formal symbol manipulation in cognition and hypothesize that consciousness is indispensable to human thought, since it is responsible for the derivation of meaning – which is what is missing in computation, as Searle has correctly pointed out.

#### **IV. Towards an Artificial Consciousness**

Having explored the philosophical foundations of artificial intelligence at some length, perhaps it is time to get back to our original concerns – where are the intelligent machines that we have read about in science fiction novels and that were once promised to us by optimistic researchers? If one were to pose this question to modern workers in the field of AI, many might reply that in some sense we have already produced intelligence, or at least certain aspects of it. There appears to be a general belief that once we have enough computing power and algorithmic complexity we will be able to engineer every characteristic of human cognition (at least as we

envision it) and thereby complete the intelligence puzzle, at which point we will be left with a genuinely cognitive creature. However, it follows from our previous discussion of computationalism that there must be something more to thought than formal symbol manipulation. So it seems that what is needed here is a more robust set of philosophical assumptions regarding the nature of mental phenomena that would allow us to make sense of the current state of AI.

Over the last two decades, the issue of consciousness has emerged as a major topic of discussion in philosophy of mind and cognitive science in general. In this paper, I will not go into detail regarding my views on the subject, as it is an extremely complex issue to address, which is why it has generated so much debate and controversy over the years. Instead, I will simply claim that consciousness is a real and exclusively mental phenomenon, which means that it is only present in creatures that have a mind. Humans are clearly such creatures, possessing the ability to think and to be conscious. Although there is no agreement on even the most basic properties of consciousness, it can be roughly described as follows: it is the subjective and ontologically first-person state of awareness that we are in when we are awake, where we experience qualitative sensations and reflect on mental objects, such as our thoughts and our own existence in the world. The true importance of consciousness is unclear at this point, but so far research has shown that it must play a central role in human-level cognition. That is why I believe we can incorporate this mysterious phenomenon into a modern view of intelligence that we can apply to problems in the field of AI.

First, let us suppose that we could resolve many of the conflicts that exist in the field today by eradicating conceptual ambiguities and formulating more precise definitions of key terms like “thinking” and “intelligence”. For instance, let us claim that the “mind” (and thus “cognition”) has two major components: (A) physical processing of information in the brain, which is akin to the symbolic computation that our generation is so familiar with. This is exemplified by subconscious human mental operations like speech production and

comprehension, acquisition of procedural knowledge (e.g., knowing how to ride a bicycle), and other processes to which we have no conscious epistemic access; and (B) consciousness, which has been described earlier, but which has the following additional responsibility – to handle issues of “meaning” in cognition, either by possessing some sort of “intrinsic intentionality” and allowing for the extraction (“derivation”, “interpretation”) of meaning from physical objects, or by somehow creating the illusion of meaning for the agent (e.g., the intuition that humans “understand” language and know the meaning of the words in a sentence, rather than simply manipulating *uninterpreted* symbols).

Then, in light of this proposition, we could claim that “thinking” (and thus “intelligence”) can be of two types as well – one that is at the level of (A), which is in principle no different from other bodily processes, like digestion, but nonetheless forms the necessary basis for the sort of cognition that humans possess; and another at the level of (B), which is characterized by a conscious and meaningful flow of mental objects in our minds (for example, it is likely that this is what Descartes’ meant when he declared “I think, therefore I am”, since there must have been a conscious thought in his mind that allowed him to produce the famous linguistic utterance – it is nearly impossible to imagine Descartes making this claim without actually being aware of himself thinking).

Now, having incorporated consciousness into a new framework for intelligence, we can make another attempt to examine the field of AI and try to understand the difficulties it has faced. What immediately becomes clear is that most (if not all) of the research up to this point has been concerned with only one type of thinking – the first one described above (this is illustrated by our pre-occupation with “information processing”, primarily in the form of digital computing). It is also evident that even if one day we will develop a purely computational solution to every “easy problem” (as David Chalmers has called them), such as language and learning, we will only succeed at replicating (A)-type intelligence. Searle would likely have no objections to this outcome, since such a system would lack intentionality and thus cannot be

called “intelligent” in the same sense of the word as applied to humans. However, it would also be foolish to halt our current efforts (as Searle would probably suggest), since this type of thinking is an important starting point for the AI endeavour and the resulting product might even manage to pass the Turing test on its own, meeting the requirements for a derivatively intelligent creature of the sort embraced by Dennett and Brooks. Therefore, it appears that current AI research will not be successful in engineering human-level intelligence until the issue of consciousness is properly addressed and until attempts are made to replicate (B)-type thinking, as it is described above. It seems that all these years we have been operating under a fundamental misconception – we thought that we wanted to create an artificial intelligence, but as it now turns out, what we really wanted was to give birth to an artificial consciousness that is akin to our own.

## **V. Concluding Remarks**

In this paper I have examined the philosophical foundations of artificial intelligence in order to gain more insight into the failures and successes of the field, and also in order to develop a more precise and constructive account of intelligence. I have had some success in analyzing the views of prominent philosophers such as Searle and Dennett and in identifying their central beliefs and arguments. I have used this knowledge to try to put forward an outline of a new theoretical framework that may be adopted by AI researchers. It consists of a dichotomy between two types of thought and an emphasis on consciousness as that which allows for a meaningful interpretation of the world. With that picture of intelligence in mind, perhaps we may someday look back in history and wonder why it has taken us so long to accomplish our age-old dream of creating cognitive beings like ourselves.

## REFERENCES

- Brooks, R. A. (1991). "Intelligence without Representation". In J. Haugeland (Ed.), *Mind Design II* (pp. 395-420). Cambridge: MIT Press (1997).
- Copeland, J. B. (2002). "The Chinese Room from a Logical Point of View". In Preston, J., & Bishop, M. (Eds.), *Views Into The Chinese Room* (pp. 109-122). Oxford and New York: Oxford University Press (2002).
- Dennett, D. C. (1981). "True Believers: The Intentional Strategy and Why It Works". In J. Haugeland (Ed.), *Mind Design II* (pp. 57-80). Cambridge: MIT Press (1997).
- Dennett, D. C., & Haugeland, J. (1987). "Intentionality". In R. L. Gregory (Ed.), *The Oxford Companion to the Mind*, Oxford University Press (1987).
- Dennett, D. C., & Searle, J. (1995). An exchange on Searle's "The Mystery of Consciousness". In *The New York Review of Books*, (Vol. 40, Num. 20, December 21, 1995).
- Searle, J. (1980). "Minds, Brains, and Programs". In J. Haugeland (Ed.), *Mind Design II* (pp. 183-204). Cambridge: MIT Press (1997).
- Turing, A. M. (1950). "Computing Machinery and Intelligence". In J. Haugeland (Ed.), *Mind Design II* (pp. 29-56). Cambridge: MIT Press (1997).