Term Paper Caleb Torres
Metaphysics
Spring 2020

Responding to Searle's Criticism of Intentionless AI

For many decades and surely into the near future, researchers have attempted to model the human brain through complex algorithms and computational models. They reason that the brain is simply a hardware device with some kind of inner program that governs behavior and thought. However, this view is not without controversy, and thinkers such as John Searle have expressed arguments against a computational theory of mind. In this paper, I wish to challenge Searle's main objections to a computational theory of mind. First, I argue that Searle fails his own position by not adequately explaining how or why a brain exhibits intentional behavior in the first place. Second, I push more on this point, and employ Dretske's framework of intentionality via his theory of representation systems to a debilitating effect. I argue that while not all programs can exhibit intentional behavior, some do, and these are the programs that Searle must ultimately address.

In this paper, I will first outline key components of Searle's arguments that I wish to address. These include his reasons why a program without an intentional system is not capable of understanding and mental processing. I also explain how Searle argues that a brain is ultimately set apart from other types of machines by virtue of its inherent intentionality.

I then offer up my two challenges to Searle's views, arguing that Searle's attacks on a computational theory of mind are not completely fatal, and that contemporary strides in AI have been enough to render certain machines with programs as exhibiting intentional behavior.

Overall, I wish to suggest that Searle's concerns regarding the question of whether a machine can think are not as critical as they seem. He admits that his "own view is that *only* a machine could think," but with the added caveat that "only very special kinds of machines," namely brains or machines with causal powers modeled after the brain (Robb 351). Thus, while Searle does not limit the prospect of thought-capable machines to being solely brains, he does little to entertain the idea of a *sufficiently* brain-like machine being capable of thought.

In his work "Minds, Brains and Programs" Searle aims to disregard a specific component of the strong AI hypothesis. Searle outlines the hypothesis as advancing the notion that AI is not merely a tool that needs leveraging, but that an "appropriately programmed computer… is a mind in the sense that computers given the right programs can be literally said to understand and have cognitive states" (332). Thus, Searle takes the strong AI hypothesis to mean that given a sufficient program or blueprint of computation, a computer can mirror the cognitive qualities of the brain. However, as we will see later, Searle does not explicitly offer what sufficient conditions must succeed in general to secure brain-like processes such as cognitive states and understanding, leaving a void to be filled by strong AI theorists.

Searle correctly identifies computationalism as central to the claims of strong AI. He describes computationalism as the theory regarding how brain processes are essentially computational processes realized in a specific hardware unit that is the brain. He mentions that the computationalist approach supposes a "level of mental operations consisting of computational processes over formal elements that constitute the essence of the mental" (342). Searle understands that with a computationalist framework, AI theorists are free to realize mental

processes in mediums other than the brain, stemming from the idea that "any computer program can be realized in different hardwares" (342).

However, Searle's scepticism about computationalism stems from its symbolic processing approach of modeling the brain. He mentions that by introducing symbols to represent objects in the world, the computer fails to regard the meaning of those symbols, and instead solely focuses on the way in which those symbols must relate to other symbols in a logical format. Searle mentions an example where "if you type into the computer '2 plus 2 equals?' it will type out 4 … but it has no idea what '4' means or that it means anything at all" (349). In other words, symbolic processing fails to incorporate meanings of symbols and leaves them uninterpreted. This is opposed to human mental states, where symbols not only serve logical roles, but semantic roles as well, and are inherent in the symbols themselves.

This distinction between formal roles, which Searle denotes as a "syntax", and interpreted meanings independent of formal roles- semantics - inspires his main argument. His Chinese room argument aims to demonstrate how pure symbol manipulation fails to grasp the whole account of an experience because it leaves out the semantic content of the input, symbolic manipulation and subsequent output.

In the Chinese room experiment, an individual is asked to sit in a room alone and receive story cards written in Chinese. They then receive a series of question cards written in Chinese, and then asked to output a set of answers to these questions in Chinese. The individual completes the task, but not without consulting a comprehensively developed aid that logically instructs the individual to match certain answers written in Chinese with the given inputs. In the end, the

individual is successful in convincing unbeknownst Chinese speakers that the system as a whole understands Chinese.

Searle questions if the individual really understands Chinese, or even if the system or some other aspect of the system can be said to understand Chinese. Searle concludes that it is obvious the individual understands no Chinese after the affair, and that it is even more absurd to conclude that any other component of the system can be said to understand Chinese as well. Ultimately, Searle offers the final conclusion that there is something about the human brain that cannot be replicated in other hardware and software systems.

Searle attributes this unique power of the human brain to intentionality. If intentionality describes the ability of the mind to have thoughts, beliefs and feelings about *something* - content about mental states - then Searle believes that in the Chinese room example, "as soon as we put something into the system which really does have intentionality, a man, and we program the man with the formal program, you can see that the formal program carries no additional intentionality" (347). In other words, if we take a "formal symbolic manipulation system" and insert a component with an intentional disposition, the vitality of the FSMS is lost and cannot account for anything above what the intentional component provides. This, according to Searle, provides evidence that without intentionality, any computational system is doomed to merely "produce the next state of the formalism when the machine is running," but entirely miss the content of what the symbols represent and mean.

Thus, Searle derives his criticism of strong AI and computationalist theories of mind by limiting intentional manifestations to brains. He argues that mental processes are restricted to

brain hardware, and thus no novel "software" implementations will ever succeed in reckoning with the root of the problem - replicating intentional structures unique to brain-like systems.

However, I have begun to develop some possible avenues for rebutting Searle's criticisms of computationalism, and how it seems to fail at accounting for intentionality. The first objection I will raise involves questioning *why* Searle uniquely attributes intentionality solely to brain systems. More specifically, I wish to exploit the fact that Searle does not answer adequately how non-brain theoretical systems with the same "causal powers" as brains can think and have mental states (351). Searle opens himself up to such speculations by emphasizing the *causal powers* of brains, and not the *brain itself*.

However, if we can conceive of systems other than brains having the capability of exhibiting mental powers, the question of how we determine the *threshold* of being a sufficiently brain-like system is critical. However, from my estimation, Searle appears to conclude that mental activities such as intentionality are only linked to "biological phenomenon" by means of a heuristic, as opposed to a case of positive evidence that actually demonstrates this thesis. He mentions that those who believe that programs could prove sufficient in producing mental activity are guilty of "abiding in dualism: the mind, they suppose, is a matter of formal processes and is independent of specific material causes" (351).

My main argument here is to show that since Searle admits brains are themselves machines, albeit a very special case, then we have grounds to 1) speculate at what threshold a machine must arrive to become sufficiently like a brain *without ultimately being a brain* and 2) claim that for brains themselves, there is still an inkling of computational structure, and thus a

program from which we may gather and rudimentary explanation of mental processes such as understanding.

Concerning the first point, Searle mentions that a brain is “a very special kind of machine” that incorporates a special set of “causal powers as brains” (351). He also claims that these qualities might arise in virtue of the brain's biological structure, claiming that “whatever intentionality is, it is a biological phenomenon” (351). However, I believe that Searle is too quick to judge here, and does not consider the possibility of a machine that is *sufficiently* biological. For example, we may consider an analogy: a smartphone and a pocket calculator are correctly considered two fundamentally different machines, but the smartphone can meet the requirements of sufficiently being a calculator with the aid of a calculator emulator application. Thus, in this case, the pocket calculator is a special machine, but the smartphone approximates this special machine, itself being a more general machine.

Hypothetically, we may conceive of a machine that is not like the brain, but sufficiently like the brain. This hypothesis may sound exhausted, especially after criticisms that have arisen such as Block’s Chinese nation argument. However, I still believe that those who argue as Block miss the point that while such a notion may sound absurd, it is nonetheless conceivable.

Thus, the question we must leave for Searle is where on this approximation does intentionality arise? Where is the threshold? Even if the threshold is infinitesimally close to being a complete brain machine, we then have a situation where a non-brain machine is capable of intentional acts, in which case there must be something about the program that is reflective of the intentional acts of the machine. Searle can no longer rely on the special status of the brain to

explain away such a program because the machine is *not a brain*, but altogether a general machine that is running a program to function.

Taking the argument even further, if we consider a complete brain machine itself, we still perceive hints of computational underpinnings at work, and observe that they are strongly correlated with mental states. This is why fields such as cognitive science and computational neuroscience are successful. That is, there seem to be correlations between our computational models and the actual states within brain processes. However, if we take Searle's account and believe that the computational models add nothing to our brain system, because the brain is only realized with its full intentional structure, these models should have nothing to offer us in terms of explanatory power. This seems off, and will only continue to sound goofy as our science augments.

My next argument considers intentionality directly, and suggests what Searle attempts to deny - that non-brain machines are capable of intentional states. Drawing inspiration from Dretske's work on representational systems, I attempt to show that computational systems are capable of having intentional stances toward external phenomena.

Dretske speaks about the intentionality of systems insightfully, concluding that "a cognitive system is a representational system of some kind, presumably of Type III" (322). A representational system of type III consists of "a power to indicate that is independent of the interests, purposes, and capacities of any other system" (313). By indication, Dretske means the ability for some event A to correctly signal some other event B. Dretske claims that in natural representation systems, an sign A indicating some external event B is not arbitrarily construed, but arises naturally through evolutionary means. Dretske also mentions that "the functions

determining what these signs represent are also independent of such extrinsic factors" (313). In other words, the functions that these indications take on are non-arbitrary as well.

Ultimately, Dretske claims that a system of representation is intentional by virtue of being independent from arbitrary signatures of representation. Unlike Dretske, Searle never discusses the sufficient conditions for intentionality, and why they ultimately rely on being realized in a biological system. This is where Dretske's account of intentionality in natural representation systems fills the void left by Searle, and to Searle's dismay, actually leaves the door open for a non-biological system to still pass as an intentional system.

I argue that while a hard-coded program that accepts inputs, follows a rule, and expresses outputs cannot possibly operate within the type III representational framework, programs that are not specifically hard-coded but nonetheless act as programs can secure a kind of intentionality. These programs are akin to machine learning programs which are prevalent in today's AI space. Much like Dretske's type III representational systems, they follow patterns only if they are deemed successful in achieving some goal, without needing explicit direction or constraints on how to go about achieving that goal.

A machine learning program, such as a reinforcement learning program, is tasked with optimizing a route toward some goal, given large sums of data and a very general algorithm that guides the program in evaluating its progress as it seeks to find some optimal solution. The program is not explicitly instructed on how to optimize its decisions based on the data, but is given a concrete method for measuring how one solution stands in relation to another. Yet ultimately, the program is left to "naturally" choose the best route of action.

Thus, we may consider a reinforcement learning program as an analog to Dretske's type III representational system. The machine derives countless indications or correlations, and forges representational functions on its own. The program is allowed to represent features of the data on its own. The machine does not arbitrarily derive representational systems either, since it must govern its behavior according to some notion of reward.

Thus, if we accept such an analog in machine learning programs, then perhaps we can entertain the prospect of a machine exhibiting some form of *crude intentionality*. It might not be at the complexity or saturation found in brains, but it would nonetheless exist enough to recognize. If this is true, Searle can no longer refute the strong AI claims that 1) programs do not understand, and 2) cannot tell us anything useful about the minds or understanding for that matter.

Thus, these two points challenge Searle's account that we cannot learn anything about the mind through theories of computation. The landscape of artificial intelligence is changing rapidly, and even paradigms of AI that existed a few years ago are now abandoned for new paradigms that inch closer to what we conventionally think of as "mental". Given the complexity of the human brain machine, there will always be a need for models to aid our understanding of mental processes, and thus a pressing need to continue in a direction that emphasizes building some kind of framework to guide us. Even if the mind is not computational in nature, Searle would agree that it is physical, and we see now that even generally physical phenomena are increasingly being modeled computationally.

On another note, Searle's Chinese room argument relies heavily on the claim that computational structures are not and never intentional structures. This is because he takes them

to have no concept of semantics, or content to the symbols they manipulate. However, I believe that this only applies to certain types of programs, and that other types of programs are actually quite intentional. In fact, AI has progressed so much that programs must be intentional if they are to succeed, since they are bereft of any instructions. The programs are only given data. Dretske is even quoted to have said regarding artificial systems that "to get genuine intelligence you need the right kind of history, the kind of history that will establish an explanatory connection between what is represented (content) and the behavior that this content helps to explain" (Dretske 216). This "right kind of history" is the often crazy data that machine learning programs receive, and yet they find ways to produce optimal solutions and actions that surprise us on a consistent basis.

Thus, to recap our discussion, I affirm that brains and mental processes are imbued with a special feature - intentionality. I even agree that intentionality is important for qualifying some system as being "mental." However, Searle does not address why this connection exists, and this is critical to the closure of his argument. If Searle cannot offer a sufficient condition upon which a system becomes intentional, then we have no choice but to speculate if a sufficiently brain-like, but ultimately non-brain machine could possess intentional features.

Furthermore, this Searle's silence on this matter leaves us to consider the thoughts of others - namely Dretske. Dretske does in fact explain the nuts and bolts of intentionality, and either "intentionally" or not opens the door for non-brain like systems to secure some sort of intentional structure. Thus, the computational theory of mind is not as doomed as Searle claims, and should at least continue entertaining the question of whether computational systems consisting of software and hardware can effectively model and explain the brain. Searle's

arguments may have seemed complete a few years ago, but the paradigm of AI is changing rapidly, and it seems too early to write off all programs as lacking in terms of modeling mental processes.

Works Used

Dretske, F. (1993). Can intelligence be artificial?. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *71*(2), 201-216."

O'Connor, Timothy & Robb, David (eds.) (2003). *Philosophy of Mind: Contemporary Readings*. Routledge.