Haverford College

The Intentionality of Machines

An Investigation into Computers' Capacity for Agential Action

Henry Woods

Senior Thesis

Professor Macbeth & Professor Sommerville

April 27, 2018

Abstract

The philosophy of Artificial Intelligence is beset by an intolerable deadlock. In an attempt to answer the question of whether computers can perform the same actions as humans, many philosophers have posited varying solutions. Of particular note are the Machine Functionalist, Biological Naturalist, and Contextualist accounts. By examining the theories that fall under these categories, it becomes plainly obvious that little progress will be made towards deciding this debate, as philosophers of Artificial Intelligence only succeed in talking past each other. The aim of this essay then is to move this debate forward by providing a test that will hopefully create some friction. In order for us to assess computers on their capacity for agency – an essential quality of the sort of actions we are evaluating computers on, such as playing and speaking – we must judge whether computers have the right sort of relationship with the world, one that is Intentional. Whereas the three major accounts delineated in this essay fail to argue against each other on some common point, we will examine how Intentionality plays a role in human action, and why it is necessary for computers to exhibit it in order for them to be doing the relevant action, and not merely mimicking a human performing that same action. This essay will develop these arguments, finally, by examining what relationship a computer would have to have with the world in order for it to be playing chess or speaking language agentially – in other words as a human does.

Acknowledgments

I am indebted to a number of people for their significant contributions to the completion of this project. I would like to extend a big thank you to my readers Professors Danielle Macbeth and Brooks Sommerville. Professor Macbeth has been integral to directing my thoughts, and providing me sources to turn to. Professor Sommerville's input was extremely valuable in that it helped me avoid impending errors in my thinking, and facilitated in focusing my ideas. Finally, I would like to thank my friends and family for the assistance and support they have provided me throughout the writing process. I should especially thank Julian Schneider for providing me with crucial edits and generally being an excellent sounding board for potential arguments.

Contents

I. Introduction

Gordon Moore, the founder of Intel, predicted in 1965 that computer performance would double roughly every two years. Though Moore was referring specifically to processing speed, he was in fact right with respect to other aspects of computing as well. Every two years not only has processing speed doubled, but also the amount of memory that can be stored in a specific amount of physical space has roughly doubled. Even still, this hardly captures the technological developments we have witnessed since Moore originally published his essay, as the general capacities of a computer have improved tantamount. However, though computers have advanced to the point that many tasks have been made almost trivial, such as calculations and communication, philosophers are still no closer to answering the major question persisting in the philosophy of Artificial Intelligence: can computers *do* human actions?

The philosophy of Artificial Intelligence is beset by the problem of figuring out whether computers can exhibit intelligent behavior. To solve this sort of problem, the Turing test was created. Functionally the test requires an evaluator to ask a series of questions to both a human and computer subject, the identities of which are unknown to the asker. The test then allows the evaluator to judge whether the answers given by the computer are indistinguishable from those produced by the human. As we will observe in a later section, the Turing test is far too broad and fails to address the question above. Many philosophers, as well as myself, echo this opinion as they are not so easily mollified by this test. The resulting philosophy that attempted to supply an answer to whether a computer can perform actions such chess playing and language speaking is especially polarized. Though the debate surrounding this question has many different positions, three primary ones are laid out in the proceeding sections. Perhaps notably, much of the work that will be cited in this essay is from some years ago, and one might rightfully be anxious that

the works and the thoughts contained within are no longer relevant. But, the core theories expressed by each of the three positions equally address modern technology as they do prior technologies, and as such still remain highly pertinent today.

Emblematic of the thorniness of the central issues of this debate is that though it has existed since the mid 1950s, minimal progress has been made towards arriving at a solution to its primary concern. What seems necessary then at this juncture is to find some common ground, some similarity between the accounts that also serves a necessary component of human action. More precisely, what this essay will attempt to do is argue for an aspect of human action that should be looked for in order to determine whether computers can do actions such as play chess or speak language in the same sort of way as a human can. To be clear, this is not an examination of *what* computers do when they appear to play chess or speak languages, rather this essay is proposing a test to see if *how* computers appear to play or speak is how humans do these same actions. Essentially, the test will allow us to establish whether computers are merely mimicking a human performing the action, or actually doing it. This test will focus on one feature of human action that differentiates human action from other sorts of actions: Intentionality. By employing a test of Intentionality, we can determine whether computers exhibit agency in the actions they perform. Davidson provides us with the definition of agency this essay will rely on:

> A reason rationalizes an action only if it leads us to see something the agent saw, or thought he saw, in his action-some feature, consequence, or aspect of the action the agent wanted, desired, prized, held dear, thought dutiful, beneficial, obligatory, or agree able…Whenever someone does something for a reason, therefore, he can be characterized as (a) having some sort of pro attitude toward actions of a certain kind, and (b) believing (or knowing, perceiving, noticing, remembering) that his action is of that kind. (Davidson 1963, 685)

Davidson's view of agential action is that it requires Intentionality.[1] Agency, an intrinsic quality of many human actions,[2] is the standard to which computers must meet if they are said to be doing the same sort of action as a human. As will be examined in two cases later in this essay, what will decide whether computers have this capacity for agency – i.e. whether they meet the necessary conditions for it – is if they have relevant Intentional states when they perform the given action.

II. Section A. Machine Functionalism

Hillary Putnam was amongst the first to consider and develop the Machine Functionalist theory. Though Putnam is the originator of the position as such, the basic idea that he proposes can be found in Thomas Hobbes' *The Leviathan.* In both, the underlying concept is that the human mind is doing nothing more than a sort of high-level calculation. Hobbes writes, "For reason, in this sense, is nothing but reckoning (that is, adding and subtracting) of the consequences of general names agreed upon for the marking and signifying of our thoughts; I say marking them, when we reckon by ourselves; and signifying, when we demonstrate or approve our reckonings to other men" (Hobbes 1998, 29). When we receive input from the world, we are in some particular mental state caused by the input received. In this mental state, we can reason about such-and-such, and then make the relevant judgments. Hobbes believes that this reasoning is really just a form of arithmetic – a belief that will be somewhat reflected in the Machine Functionalist position that Putnam and others advance.

---

[1] Intentionality will be defined later in section V.

[2] Importantly, the sorts of actions denoted in this essay have the quality of agency in human action. Specifically, this essay looks to the actions of playing and speaking, both of which are Intentionally done when performed by a human. I am not making the claim that *all* human action is agential, just that these particular actions are.

The basic argument for Machine Functionalism hinges upon a comparison between the Turing machine, essentially a mathematic abstraction of a computer I will elucidate momentarily, and humans. Putnam believes that all the processes that a Turing machine does are, in a certain sense, also realized in the functions of a human. In order to grasp his argument, a better understanding the Turing machine is needed. Putnam describes Turing machines as follows:

> Briefly, a Turing machine is a device with a finite number of internal configuration, each of which involves the machine's being in one of a finite number of *states*, and the machine's scanning a tape on which certain symbols appear. The machines tape is divided into separate squares… on each of which a symbol (from a fixed finite alphabet) may be printed. Also, the machine has a "scanner" which "scans" one square of the tape at a time. Finally, the machine has a *printing mechanism* which may erase the symbol which appears on the square being scanned and print some other symbol (from the machine's alphabet) on that square. Any Turing machine is completely described by a *machine table*… A "machine table" *describes* a machine if the machine has internal states corresponding to the columns of the table, and if it "obeys" the instructions in the table in the following sense: when it is scanning a square on which symbol s_1 appears and it is in, say, state B, that it carries out the "instruction" in the appropriate row and column of the table. (Putnam 1960, 22)

Putnam describes the theoretical model of something that takes an input, changes states based on that input given the rules codifed in the machine table, and then provides some output based on the state that the machine is now in (the state change being "caused" by the input according to the rules). Turing machines are a particularly convenient model insofar as any computer can be theoretically reduced to a Turing machine. However, at first blush it may be mystifying as to how we get from this abstraction of a computer to a point where we witness a resemblance with humans. Evidently, the output of a Turing machine need not be just symbols printed on an output tape. Instead, Putnam asks us to consider a machine that could be modified such that it has appendages that move in response to particular inputs. The machine could be, under Putnam's theoretical framework, further modified such that all human outputs, including speech and other sorts of movement, have been realized. Furthermore, the machine could take any number of

inputs – for example, those provided by sight, hearing and so on. However, what remains less clear is how the machine also represents mental activity that is similar to a human's.

To show how one might conceive of a computer's mind as a human's, Putnam takes up a number of potential objections to Machine Functionalism in his essays "Minds and Machines" and "The Nature of Mental States." Of these objections, let us consider the concern that a machine could not know its present state. Given that a human may be able to say "I am sad" when feeling a certain way, the conceivable objection against Functionalism may be formulated as "how might a machine ascertain that it is in state A, where state A corresponds to feeling sad". Consider then a machine that receives input X, and X causes the machine to transfer through states D, C, B, ending finally in state A. In state A, the machine may have a rule that when requested, causes the machine to output the statement "I am sad". Now, if human subject Jones were to be in such a mental state that he feels sad, he may utter the statement "I am sad". In this way, one should be able to say of Jones that his being in a mental state of sadness is what *caused* him to utter the phrase "I am sad". Putnam believes this to be sufficient evidence to show that the following statement formulation, 1 and 2, should be acceptably equivalent:[3]

1) The machine was in state A, and this caused it to print: "I am in state A"
2) Jones was sad and this caused him to say "I am sad"

This may lead to clarifying questions of how the state was ascertained – comparable perhaps to how Jones knew he was sad – but these questions only mislead. Putnam explains arguing, "the traditional epistemological answer … - namely, "by introspection" – is false to the facts of this case, since it clearly implies the occurrence of a mental event (the "act" of introspection) distinct from the feeling of [sadness]" (Putnam 1960, 24).

[3] I do not hold these assertions to necessarily be equivalent, rather machine functionalists believe the causal properties of both the machine state and pain state of Jones are as both states seem to cause a similar output.

Importantly for our investigation, the Machine Functionalist account incorporates the concept of Intentionality. Daniel Dennett explains how the account may make this inclusion in his essay "Intentional Systems." For some machines, one may be able to ascribe its behavior to particular design assumptions. That it is to say, one may make the prediction that, say, a soda machine will dispense a soda when a particular amount of money is inserted because that is what the machine is designed to do. This sort of prediction can be labeled the design prediction. Similarly, there is the physical prediction that something will perform a particular way due to the laws of nature. However, Dennett holds that neither sort of prediction will suffice if trying to predict a modern chess computer's next move. Rather, Dennett proposes an alternative means of prediction: "A man's best hope of defeating such a machine in a chess match is to predict its responses by figuring out as best he can what the best or most rational move would be, given the rules and goals of chess" (Dennett 1971, 89). Dennett suggests that one must consider the chess machine as a rational agent in order to beat it as mere reflecting on the design of the software is too complex of a task to parse out what the next move will be. If it is indeed the case that one must predict moves in the way Dennett describes in order to beat a computer at chess, then Dennett believes this conclusion follows:

> This third stance, with its assumption of rationality, is the Intentional stance; the predictions one makes from it are Intentional predictions; one is viewing the computer as an Intentional system. One predicts behavior in such a case by ascribing to the system the possession of certain information and by supposing it to be directed by certain goals, and then by working out the most reasonable or appropriate action on the basis of these ascriptions and suppositions. (Dennett 1971, 90)

By treating the machine as an Intentional system, it is certainly less absurd to presume that the information the computer possesses is in the form of beliefs and desires. This is especially true when considered in terms of Dennett's framework of Intentional systems, that an Intentional system need not really have beliefs or desires, but can merely be ascribed these Intentional

idioms. As such, an appropriately programmed computer can be said to be an Intentional system in the same way a mouse can be said to be an Intentional system in its desiring to eat cheese or avoid being killed. It is in the ascription of Intentionality to these systems that the system appears to us as an Intentional system. Dennett's Intentional stance thusly avoids cluttering the concept of Intentionality with metaphysical difficulties.

III. Section B. Biological Naturalism

John Searle proposes the Chinese Room as a counter-argument to claims made about strong AI. Strong AI is artificial intelligence that stands directly opposed to weak AI and closely resembles the Machine Functionalist position presented in the section above. Generally speaking, claims from the perspective of weak AI are that computers are just very powerful tools. What AI will allow us to do is perform certain tasks, such as testing hypotheses, in a more precise way. However, Searle is not concerned with the claims made of weak AI, taking issue in his essay with those made of strong AI. Searle summarizes these claims as follows:

> But according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states. In strong AI, because the programmed computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations. (Searle 1980, 417)

Strong AI then is not arguing that the computer is merely a very effective tool, but rather that the program or computer itself understands a problem that is given to it. This is a constitutively different claim than those made of weak AI. For those supporting strong AI, they argue that what the computer is able to do is *exactly* the same as what a human is doing. While not being a physical brain it is, in essence, a mind. This artificial instantiation of a mind is therefore able to process information and provide output that seemingly demonstrates that the computer had

thoughts about the problem, or whatever the input, presented to it. Weak AI, however, merely maintains that when a computer provides an answer, the answer is the result of some user directed process. The user gives the computer some specific task and using a given set of information, the computer is able to quickly, and accurately, provide an answer. In this second case, no understanding of a problem is demonstrated, what is represented is that the computer was able to successfully complete a menial task quickly, and perhaps more efficiently than a human. Searle is skeptical of the strong AI position, as he believes strong AI asserts two contentious points:

1) The machine literally understands problems and information provided to it, and so the answers it provides are derived from *this* understanding.
2) It could be said that the machine actually explains a humans ability to understand a story and answer questions about it. (Searle 1980, 417)

With the strong AI position now being laid out Searle moves on to consider the Chinese Room, a thought-experiment that aims to provide an argument against strong AI. The premise of the experiment is that a person named Julia[4] is locked in a room. Julia is a native English speaker who has no understanding of Chinese whatsoever – she may not have ever seen a Chinese character before. Julia is then given two stacks of paper containing Chinese writing, and one stack of paper with English writing. The stack containing English writing has rules, that Julia can plainly understand, which allow her to correlate certain elements from the second stack to the first stack. However, unknown to Julia is that the stacks with Chinese characters correspond to a story written in Chinese and questions about the story. Julia's task within the room is to use the rule stack provided to correlate symbols in the first two stacks, and then output these comparisons – effectively from an outside perspective answering the questions to the story. Over time, as Julia gets more practice at applying the rules to the Chinese stacks, and as the "scripts"

[4] Searle uses himself in the experiment, but for the sake of my essay I'm changing the name since using Searle as the subject of the experiment might confuse rather than help with explaining the experiment.

being given to her better help her correlate symbols, her answers should accordingly become increasingly more accurate. That is to say, eventually the answers provided to the Chinese stories and questions should be indiscernible by a native speaker from the answers Julia could provide to English equivalents. Searle contends now that by extending the claims of strong AI proponents to the Chinese room (as the thought experiment is supposed to simulate a computer), we would expect them to argue (1) Julia literally understands the Chinese stories being given to her. Furthermore, (2) Julia's understanding of the story explains human understanding, as both humans and computers understand in the same way. Searle believes these two hypothetical assertions to be false.

Searle responds to putative assertions (1) and (2) as follows: "it seems to me quite obvious in the example that I do not understand a word of the Chinese stories. I have in-puts and outputs that are indistinguishable from those of the native Chinese speaker, and I can have any formal program you like, but I still understand nothing" (Searle 1980, 418). Searle goes on to emphasize that the machine instantiation does not demonstrate sufficient conditions for understanding. This is because, Searle points out, that what is actually occurring in the Chinese room is mere symbol manipulation. The Biological Naturalist accounts believes this poses a problem for strong AI supporters since symbol manipulation does not allow for understanding as the process lacks Intentionality. Though Julia is a human, and as such necessarily enjoys Intentional states when performing relevant actions, when she is asked to execute the task in the thought experiment she lacks Intentional mental states. Analogously, an electric kettle turns off once water is boiling, but it is not as if the kettle is directed at the state of affairs that constitutes boiling water. The reaction to a conditional being satisfied – the water reaching a temperature of 212 Fahrenheit – does not demonstrate that the kettle has understanding and, correlatively,

Intentionality. For comparison, if a human were to turn off a kettle, there is inherently Intentionality as the human likely has some belief about the water, that is to say she is directed at the world in the right sort of way.

Searle leaves us with both a positive and negative account of strong AI. The positive account is that a computer can think, but only computers that are like brains in that they have the same causal powers as brains. However, here lies the problem with AI since AI is not really about machines at all. What AI is about is programs and programs are not machines. As such, we get the negative account, which is that programs do not exhibit Intentionality as Intentionality is necessarily a biological phenomenon. Searle motions to what causes Intentionality through a hand-wavy gesture stating, "Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena" (Searle 1980, 424). Though Searle does not commit himself to any particular aspect of the brain that allows for humans to have Intentional states, he maintains that it must be biologically dependent in the same way many other biological phenomena are. It would be ludicrous, Searle suggests, to judge that a computer program could simulate the production of sugar - the result of photosynthesis. Though Searle believes that many would not try to argue that these sorts of phenomena could be computer-generated, due to a strong attachment to dualism some are inclined to believe that the Intentionality of the mind is not dependent on some biochemistry.[5] However, as Intentionality is like any of these other phenomena, this dependence on biological matter precludes instantiations of programs from ever exhibiting Intentionality, as programs alone cannot be sufficient for Intentionality.

[5] Though I do not really wish to go much into this dualism, essentially what this dualism argues for is that the mind is an independent, formal process that is not dependent on causally related material, namely the brain. Effectively, the brain and mind are separate entities.

In order to further substantiate Searle's negative account, we can look to claims made by Christopher Maloney in his essay "The Right Stuff." Maloney supports the Biological Naturalist position by first discussing a common, yet erroneous refutation of a materialistic identification of mental processes. The objection is that what the materialist (naturalist) believes is missing for subject Julia, or some theoretical computer instantiation of Julia, is "some inscrutable stuff, some secret, slimy secretion" (Maloney 1987, 367). Believing this to be the position of the materialist, the objector might, mistakenly, make the case that one could conceive of some extraterrestrial silicone based life form that has the capacity to understand Chinese, write poems and so on. As such, a neuroscience that is partial to carbon-based life would therefore fail to identify whatever the "secret stuff" is that is intelligence. Maloney argues that this is wrongly used as a refutation of the Biological Naturalist position by contending:

> Its error is evident once we realize that we can also imagine that, possibly, wherever in the universe understanding abounds it is *essentially* organic. Does this show that it is possible that intelligence is necessarily organic, and given at least one plausible version of modal logic, that intelligence is therefore necessarily organic? This way of arguing about prospects of a science is absurd, for it rests on imagination indiscriminately fuelled by dogma and insight. Surely, though we can imagine that possibly water is not H_2O, it remains that chemistry rightly takes water necessarily to be composed of hydrogen and oxygen. Nothing else could be water, regardless of what we fancy. (Maloney 1987, 367)

Maloney defends Searle's negative account by effectively reversing the aforementioned argument. What Maloney contends is that this counter-argument really demonstrates that intelligence is essentially organic – a point that favors the materialist account. Rather than arguing that intelligence and essentially human action – identified by the presence of Intentionality – must necessarily require a biological organism composed of carbon, Maloney shows that the Biological Naturalists like him and Searle are only positing that the "right stuff", as it were, is something organic. The Intentionality both Searle and Maloney are concerned with requires some biological, that is to say organic, material in order to exist, otherwise a program or

other artificial structure cannot be said to be performing anything other than mere symbol manipulation.

IV. Section C. Contextualism

The Contextualist position relies heavily on the work of Merleau-Ponty and Heidegger as it relates to perception and its role in understanding and engaging with the world. This is due to the influence these philosophers had on Hubert Dreyfus and his book *What Computers Can't Do*. In the book, Dreyfus explains that it is the human faculty of perception that allows humans to perform certain actions and be an agent in the world. It is through experience and involvement with the world that provides us knowledge of it. This is problematic for computers, Dreyfus will go on to argue, because they can only be provided with explicit rules for performing a certain action – discounting a computer from ever being able to match a human's ability in certain actions. In order to grasp Dreyfus' argument though it may be best to start with Merleau-Ponty's account of attaining knowledge of the world. Merleau-Ponty writes, "In Müller-Lyer's illusion, one of the lines ceases to be equal to the other without becoming 'unequal': it becomes 'different'. That is to say, an isolated, objective line, and the same line taken in a figure, cease to be, for perception, 'the same' (Merleau-Ponty 1962, 13). In experiencing some visual stimuli, say the Müeller-Lyer lines, one may at first perceive them as being unequal. However, through visual perception we are also able to perceive them as being equal by directing one's focus to certain parts of the line such that it no longer appears to one as being unequal.

This sort of perceptual capacity that humans possess that allows us to take in the world and engage with it is what differentiates the way humans play chess from the way computers do. That is, through a visual perception of the board in chess, human chess masters are able to make

intelligent moves without needing to calculate thousands of moves ahead. Dreyfus argues that herein lies a difference between computers and human chess players. The computer, using explicit pre-programmed rules, calculates some 26,000 different moves. Human chess masters, however, only need consider 100 or 200 alternatives when deciding a particular move. Dreyfus argues the cause of this is that "Once one is familiar with a house, for example, to him the front looks thicker than a façade, because he is marginally aware of the house behind. Similarly, in chess, cues from all over the board, while remaining on the fringes of consciousness, draw attention to certain sectors by making them appear promising, dangerous, or simply worth looking into" (Dreyfus 1984, 16). Dreyfus also points to studies conducted by De Groot that further demonstrates that superior perceptual abilities were an important factor for talented players. The ability to perceive the board as such is not possible for a computer to acquire. This is because through experience with chess the human player learns to attune her attention to certain parts of the board that may often lead to a stronger move. The computer, however, must explicitly calculate out potential moves before deciding.[6]

The difference becomes plainly obvious for Dreyfus in the case of language acquisition. One such difficulty arises in considering how humans perform information processing. When hearing an utterance, a human is capable of parsing out the ambiguity, which is to say a human uses context to remove the extraneous parts of the sentence and to arrive at its meaning. A computer, however, must be programmed such that it is told definitively which parts of the sentence are stressed or inessential and so on. Dreyfus believes that the only way in which a computer may not require explicit rules already programmed into its software instructing the

[6] Dreyfus exposes himself here to an objection that will be considered later, and indeed is necessary for our discussion. Dreyfus essentially argues that computers and humans do not do the same actions since they do not do it in the same way, but perhaps it is possible for computers to perform the same actions as a human but in a different way.

program as to what specific parts of the sentence to focus on would be to have the machine "learn" languages – precisely to have a child's ability to learn. However, this response fails to get at the root problem of artificial intelligence, as Dreyfus believes the issue to be in *how* a computer's learning would be different than a human's. This is to say the presupposition that all that is lacking is some ability to learn is to merely believe that learning a new word is some conditioned reflex of associating words with objects. On occasion, a child may learn a word when an adult points at an object and says what it is. But, Dreyfus does not believe that is sufficient for learning a language and uses Wittgenstein to make this point:

> But Wittgenstein points out that if we simply point at a table, for example, and say "brown," a child will not know if brown is the color, the size, or the shape of the table, the kind of object, or the proper name of the object. If the child already uses language, we can say that we are pointing out the color; but if he doesn't already use language, how do we ever get off the ground? Wittgenstein suggests that the child must be engaged in a "form of life" in which he shares at least some of the goals and interests of the teacher, so that the activity at hand helps to delimit the possible reference of the words used. (Dreyfus 1984, 22)

Dreyfus does not believe that language acquisition is purely the affixing of some label to a particular, delineated concept or object. There is some element essentially missing in this formula – there is something particular about humans that allow them to acquire a language in some way other than in this "point and repeat" method. AI researchers, however, seem to believe that language may be learned in just this way. Dreyfus explains stating:

> For the AI researcher it seems to justify the assumption that intelligent behavior can be produced by passively receiving data and then running through the calculations necessary to describe the objective competence. But, as we have seen, being embodied creates a second possibility. The body contributes three functions not present, and not as yet conceived in digital computer programs: (1) the inner horizon, that is, the partially indeterminate, predelineated anticipation of partially indeterminate data (this does not mean the anticipation of some completely determinate alternatives, or the anticipation of completely unspecified alternatives, which would be the only possible digital implementation); (2) the global character of this anticipation which determines the meaning of the details it assimilates and is determined by them; (3) the transferability of

> this anticipation from one sense modality and one organ of action to another. All these are included in the general human ability to acquire bodily skills. (Dreyfus 1984, 167)

What appears to be missing for the computer is a means of acquiring bodily skills, the sort of expert knowledge-how that humans have the capacity to develop.

Dreyfus argues that it is due to humans' embodiment that they are capable of learning languages and performing skilled action. By skilled action Dreyfus is referring to the sort of action that a driver or perhaps a skilled typist exhibits. When a particular bodily skill is learned, and then practiced to perfection, the action becomes consciously effortless, requiring minimal attention. Dreyfus explains this process as follows, "Generally, in acquiring a skill… at first we must slowly, awkwardly, and consciously follow the rules. But then there comes a moment when we finally transfer control to the body. At this point we do not seem to be simply dropping these same rigid rules into unconsciousness" (Dreyfus 1984, 767). A skilled action is first learned by rote, which is the effortful memorization and attention that comes with following prescriptive rules. In the case of driving this may come out as directing one's attention to the mirrors, or in the case of typing, as the memorizing of key locations and staring at one's fingers while writing. However, over time each specific action takes on a degree of smoothness and effortlessness as it is brought into a collective of bodily movements. There is no longer the need to direct oneself at a particular motor action, such as a precise key press or mirror check. Rather, one is merely engaged with an action undivided, a sort of muscular gestalt.

The phenomenology of skilled action though is only realized in humans, Dreyfus argues, because of our embodied-ness. Humans are capable, for instance, of determining what action may be required before it is compulsory. A skilled driver has the ability to adjust to a hazard immediately simply through a sort of perceptual awareness. That is to say a driver may not have even realized that a car has suddenly stopped ahead of her before her perceptual awareness of the

situation has caused her foot to press down on the brake. Dreyfus believes that for a computer, this sort of anticipation is not possible. There is some prescriptive rule that must be implemented before a driving program is able to assess a situation and respond correctly. The programmer must have written in a set of instructions that tells the program to brake when there is another car that has stopped within 30 feet – in logical terms, the satisfaction of a prescriptive conditional must be met. As a human does not require this rule-based response to a given situation, and as a human is capable of anticipating without meditating on, or considering, the correct reaction, Dreyfus' account concludes that computers are unable to perform the same sorts of actions as a human, since they cannot do the actions in the same way. Computers lack embodiment, and so cannot learn language or perform skilled actions like chess or driving as a human does. By not being engaged with the world in the sort of way required for these actions, Dreyfus effectively judges computers to be non-agential.

V. A Synthesis: Intentionality

The challenge we now face is how these accounts may contribute to solving the debate in which they are situated. This will prove particularly challenging, as each account appears to be concerned with a different issue that AI faces or manages to overcome. In other words, for us to determine whether computers can ever be agential actors from the accounts considered above, we must find some common ground from which each of these positions can be approached. At our present juncture, one may decide the question for oneself by simply subscribing to one of either Functionalism, Naturalism or Contextualism. But, this solution seems unsatisfactory. To better explain why, we must reflect on what differentiates each account from the others, as each

position emphasizes some supposedly important aspect of humans or computers that is believed to be the deciding factor for whether a computer can exhibit agency.

Beginning first with Machine Functionalism, philosophers such as Putnam and Dennett are primarily concerned with causally related mental and machine states. Along with other philosophers aligned with Functionalism, they argue that the reason computers can *do* the same sorts of actions as a human is because they perform them in the same way. The way a human mind works is very similar to, indeed almost symmetrically realizable as, a machine state table. The argument goes on to claim that a Turing machine, a theoretical instantiation of a computer, could be conceived that almost exactly replicates a human performing a particular action. This is made possible due to the causal relations of mind and machine states, which the Functionalists argue is in part proof of a computer's capacity for human action. Dennett's Intentional stance supplements this claim by arguing that so long as a system can be analyzed and interacted with under the presumption of Intentionality, then the mind and computer are not incommensurable since both appear to be Intentional systems.

In turning to the Biological Naturalist position, we are confronted by an entirely different concern. Searle does not respond directly to the Machine Functionalists, which might have been achieved by arguing that minds and machines do not function in some causally related way. Rather, Searle focuses instead on the aspect he believes is most important in determining a computer's capacity for human action. His concern is that computers can never *understand* the activity they are performing, and as such cannot be doing the same thing as a human. Necessary for understanding, Searle posits, is that there must be Intentionality. This requirement is problematic for computers because he thinks only a brain can exhibit Intentionality due to whatever biological matter endows brains with this unique ability. In this way, Searle highlights

a biological consideration that the Machine Functionalists are not interested in, thereby ensuring that these two accounts fail to gain any traction with one another.

This failure to find friction that would help towards solving the debate is also apparent when considering the Contextualist account. Dreyfus does not address the issues brought up by either Machine Functionalism or Biological Naturalism, suggesting instead that the decider is embodiment and a perceptual capacity. In other words, Drefyus believes that much of human action is made possible by humans' ability to take in the world around them – an awareness of a whole state of affairs. Humans can take in the full context of their surroundings, which influences actions such as chess playing and language speaking. Humans' embodiment directs humans towards relevant referents when learning a language, or particular parts of a chessboard when playing chess. This embodiment is a necessary condition for perception, which is what allows humans to attain and perfect these skills. Of course, it is plainly obvious that perceptual ability is not a condition believed necessary by the Machine Functionalists or Biological Naturalists. As such, Dreyfus has simply complicated an already complex issue with yet another consideration that must be contemplated.

These different focuses create an intolerable divide that may only be bridged once some common ground between the positions is found. However, though each of these accounts has seemingly only complicated this debate by adding disparate concerns that effectively talk past each other, there is some commonality that can be parsed out from them as well. By pulling out this cohesion a bridge may be formed that will allow for productive dialogue, moving us towards addressing the issue we are trying to solve. The key to the synthesis of these three positions, I believe, is Intentionality. Intentionality has explicitly appeared twice so far in the accounts outlined above. Dennett and Searle, representing the Machine Functionalist and Biological

Naturalist accounts respectively, both appear to hold that Intentionality is a necessary condition computers must satisfy in order to be considered agential actors. However, though we have already encountered this term, it has not yet been specifically defined. In order for us to have a clearer idea of what Intentionality means, we should turn to its original definition as posited by Franz Brentano:

> Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object (which is not to be understood here as meaning a thing), or immanent objectivity. Every mental phenomenon includes something as object within itself, although they do not all do so in the same way. In presentation something is presented, in judgement something is affirmed or denied, in love loved, in hate hated, in desire desired and so on. (Brentano 2015, 68)

Brentano then, takes Intentionality to mean the mind's capacity to be about or directed at some property or state of affairs in the world. Intentionality characterizes mental phenomena as the mind is directed at, in a particular way, some object within itself. Intentionality is, therefore, intrinsic to human action since all mental phenomena exhibit Intentionality. As this is a classification of human mental phenomena, and is thus present when humans are perceiving, believing or desiring, computers too *must* exhibit Intentionality for them achieve a capacity to perform the same sorts of actions as humans.

Though we have somewhat delineated Intentionality, it is perhaps not yet clear why Intentionality is important in solving this debate. We can better understand its relevance by turning to Davidson's swampman thought experiment. The experiment was employed in response to Putnam's Twin-Earth thought experiment[7] as a means of examining what is essential for the existence of psychological states like knowing the meaning of a word or having thoughts. Davidson's thought experiment supposes that Davidson is standing in a swamp next to a tree.

[7] This thought experiment can be found in Hilary Putnam's "Meaning and Reference" (1973).

This tree is struck by lightning, disintegrating Davidson and simultaneously forming from the swamp muck doppelganger Davidson. This doppelganger Davidson appears to act the same way as original Davidson, speaking to friends and writing articles with no discernable difference. Davidson, however, argues that it is not the case that there is no difference stating:

> But there is a difference. My replica can't recognize my friends; it can't recognize anything, since it never cognized anything in the first place. It can't know my friends' names (though of course it seems to), it can't remember my house. It can't mean what I do by the word 'house', for example, since the sound 'house' it makes was not learned in a context that would give it the right meaning--or any meaning at all. Indeed, I don't see how my replica can be said to mean anything by the sounds it makes, nor to have any thoughts. (Davidson 1987, 444)

Davidson provides us with a counterfactual that motivates the need for Intentionality in agential systems. The swampman depicted in the thought experiment evidently has no Intentionality. Though it performs actions as if it did, which is to say, it seems to perceive people and converse with them, the swampman's mental phenomena are not directed at anything at all. In order for the swampman to be an Intentional system, its mental states must be about the world in a certain sort of way, which Davidson argues it clearly is not. When the swampman perceives his, i.e. original Davidson's house, his state of perception is not about the house in the same sort of way Davidson's perception of his house would be. As this counterfactual suggests, the molecule for molecule recreation of Davidson does not have Intentional mental states as Davidson does, thereby rendering it non-agential in the way actual Davidson intrinsically is.[8]

Though Davidson's swampman should suggest a need for Intentionality, this claim that Intentionality is necessary for agency is also supported in all the accounts mentioned previously. As was briefly touched on, Searle and Dennett argue that Intentionality must be exhibited by a

[8] One might be tempted to disagree, in the vein of materialism, with both Davidson and my assessment of the swampman. A Searlean type response would probably suggest that both entities, Davidson and swampman, have Intentional states as both have the same sort of brain activity i.e. made of the same biological material. However, as I bring out later in this section, Intentionality is more than just a certain sort of brain activity, but is also a relationship to the world in a particular sort of way.

computer for it to be performing human actions. For Dennett, this is apparent in his portrayal of the Intentional stance, whereas Searle argues this by demonstrating that Julia in the Chinese room lacks Intentionality. Dreyfus, though not arguing for Intentionality explicitly, would likely agree that Intentionality is necessary for the sort of perceptual capacities he believes humans are capable of that computers are not. It seems likely then that we can extricate some impression of Intentionality from his account as well. Dreyfus motivates how computers and humans direct themselves at the actions they perform. Humans perceive a whole state of affairs when performing an action. This sort of perception, the attending to a whole state of affairs – the complete engagement with the world – is something Searle claims is characteristic of Intentional visual experiences, "The visual experience I will say does not just represent the state of affairs perceived; rather, when satisfied it gives us direct access to it, and in that sense is a presentation of that state of affairs" (Searle 2004, 46). As such, Dreyfus too is arguing Intentionality must be present for agency as humans are the only systems that can fully engage with the full state of affairs that the world presents to them.

Intentionality certainly appears to be a way to help settle the issue we presently face. Humans are Intentional systems, as Davidson demonstrates, while Dennett, Searle and Dreyfus concur. As such, determining whether computers are Intentional systems would help decide the question of whether a computer can exhibit agency. Indeed, it is altogether possible to conclude that it is a *necessary condition* that an agent be an Intentional system, given humans as our paradigm of agency. Therefore, if Intentionality can be thought of as a necessary (not sufficient) condition for assessing the agential nature of computers, then one of two things can be taken from the proceeding cases. Either, (1) the situation discussed demonstrates Intentionality in computers, thereby satisfying a necessary condition of an agential system. This does not,

however, prove definitively that a computer can perform human action – it merely preserves the potential of computers to perform human actions. Or, (2) if a particular case indicates that a computer does not, in fact, demonstrate Intentionality when performing the relevant action then it should be said that in *that case* a computer is not doing the appropriate action to be considered an agent. What I mean by this is that if one believes that the computer is not exhibiting Intentionality when "playing" chess then the computer is merely *mimicking* "playing" chess – it is not phenomenally "playing".

But, this does not solve the issue of how to determine whether a system is Intentional. Though Intentionality provides us a mark for whether or not a system is agential, it remains somewhat unclear what it would mean for a system to exhibit Intentional mental states. There are two ways we may approach this. The first of these is to take Dennett's Intentional stance and apply it to the two cases that will be examined below. If the case in question *appears* to show computers exhibiting Intentional states, that is one could ascribe desires, perceptions and beliefs to the system, then the system is proven to be Intentional. However, this does not seem satisfactory. As will be discussed later, little is accomplished by attributing these states to a particular system, which leads us to the second and more appealing option. Rather than merely attributing Intentional states to systems, something more rigorous is required that will allow for a way to plainly determine whether computers have these states. In order to do this, we must consider how these states relate to the world. Intentional states have two directions of fit – world-to-mind and mind-to-world. States (such as desires) have world-to-mind direction of fit, as it is such that the world should change in order to match the mental state in question. For instance, if one has a desire to be a millionaire, and one is not a millionaire, the world should change in some way so as to accommodate this desire.

Just as Intentional states with world-to-mind direction of fit can be at odds with the world, so too can Intentional states with mind-to-world direction of fit. One can have a fallacious belief, for instance, if one believes it to be raining outside and it is not in fact raining. Similarly, one can misperceive a small dog as a cat, and it is one's perception that is wrong, not the world. In this way, through the cases of chess playing and language speaking, we may test to see whether computers can have mental states that do not necessarily align with the world. An important aspect of Intentional states, both those with mind-to-world and world-to-mind directions of fit, is that they can fail to capture something in the world. Whether one desires something that is not yet realized, or believes something that is not true of the world, in such instances there is misalignment between the agent and the world. This brings out a unique sort of relationship between agent and the world, one that seems to uniquely demonstrate an important characteristic of Intentionality. The consideration then that should be made in the cases examined below is whether computers are related to the world in the sort of way humans are: that they have states that may be at odds with the world their mental states are about. One should not be, as one may be tempted to do, examining the cases in order to establish if computers do the same actions as humans *in the same way*. Rather, one should judge computers to either be doing or mimicking a particular action by whether they have the same relationship with the world as humans do when they perform the relevant action, i.e. if they are doing an action Intentionally. What is more, I leave the judgments to the reader or for subsequent work, as this essay wishes only to demonstrate the way in which one would go about assessing computers for agency.

VI. Case I: Chess Playing

Chess provides an excellent example of strategy mastery and an ability to plan ahead. It is for this reason, as well as the fact that chess-playing computers have been around for some time,[9] and are frequently referred to in the philosophy of artificial intelligence, that we will investigate a computer's ability to "play"[10] chess. A human player expertly plays chess in one of two ways. Either, this playing is done the way that Dreyfus brings out – the chess master is perceptually aware of the whole state of affairs of the board, and as such is able to focus her attention only on particular pieces of importance. For these select pieces, the chess master may calculate potential moves and assess the strength of each. This highly complex calculation is made possible by the player's contextual awareness, the state of the board and the pieces that would lead to the strongest moves; an awareness informed by her perceptual faculties, as well as practice and experience. It may even be that, phenomenally, the chess master is not aware of the complex calculations occurring as she examines the board. Alternatively, one may believe that the chess master is just very good at performing algorithmic calculations that allow her to contemplate many turns ahead of the present situation. However, in either instance one thing remains true: the expert player is Intentionally playing the game. She has certain Intentional states while playing the game, which we know to be true as humans are inherently Intentional systems with such states. Importantly though, we should observe how the chess masters' Intentional states can fail to capture some state-of-affairs in the world. An obvious instance of this might be that she has beliefs that move X is the strongest move she could make, when it would have been more

[9] Deep-blue first beat a human world champion in 1996.

[10] I use the word "play" in quotation marks because I do not wish to imply that computers necessarily are "playing" these games. For the sake of clarity though, I will not do this throughout this section as it may become confusing – just realize that whenever the term "play" with or without quotation marks appears it is not a judgment of the action being performed.

expedient for her to make move Y. Incorrect beliefs such as this demonstrate the way in which a computer should be related to the world if it is to be truly playing chess.

When playing a highly advanced chess computer, we are subject to ascribe Intentional states to the machine in the way that Dennett brings out in his Intentional stance. The computer opponent appears to desire to win the game, to perceive the particular piece it is moving and to believe that the move it performed was the strongest. Indeed, there even exist computer programs that "learn" the game through "playing." The computer, after being supplied the rules required for chess, can become better at winning by simply storing information from games "played". But, here lies the issue with the Intentional stance: Though we can certainly say that the computer appears to have such and such a state, that it believes X or desires Y, this says nothing of whether it is related to the world in the same way a human player is positioned when playing the game. Effectively, nothing can be said about whether what appears to be an Intentional state really is just that, an Intentional state. What decides whether artificial intelligence has the possibility of agential action is only its relationship to the world, not whether it performs the action in the same way or even if it appears to be Intentionally playing chess.

By merely focusing on what appears to be occurring when the computer plays the game, the question of whether the computer is playing or mimicking a human cannot be decided. If a computer can only be said to desire to win the game, then it seem ludicrous to conclude that it must be actually desiring to win the game. This is because, as was brought out in the previous section, the relationship that the agent has to the game must be one in which a state can be at odds with the world. A human player may be wrong about a particular belief, or may fail to win a game. To demonstrate, take for instance a human that was just randomly moving pieces around the board. Perhaps even this human knows how to play chess, but his mind is focused on

something else and he moves one of his pieces merely in response to his opponent completing a move. Suppose even further that these moves are legal moves, and for all intents and purposes a legal game of chess is being playing out. However, would one conclude that he, our subject player, is actually playing chess? This case seems very similar to Davidson's swampman, in that this person has the capacity to have Intentional states and to be directed at the board in an Intentional manner, but for whatever reason is not attuned to the game in the same way that say his opponent is. If one wished to draw out this experiment, perhaps our subject is in fact an exact clone of his opponent. So long as one believes that swampman is not Intentional, the subject of this experiment should be thought of as similarly lacking Intentional states that are *directed at the game.* Our clone player clearly does not desire to win the game, as how could he when his mental states are not directed at the game he is playing? Though moving pieces randomly and without strategy does not alone preclude the clone player from desiring to win the game – as it is possible for one to not be aware of the rules of a game yet still desire to win they are playing – it is the fact that the clone's mental states are not directed at the game he appears to be playing that does so. We may ask similar questions as we did above of other Intentional states, such as believing and perceiving states.

It is therefore not enough to simply ascribe Intentional states to a computer, as even "players" with the capacity for Intentionality need not be Intentionally directed at the game they are appearing to play. Computers must be Intentionally directed at the game in such a way that its states can be about states of affairs that are at odds with the world. For a computer to be an agential player, it may have, for instance, a desire to win the game. This state has the makings of an Intentional state if the desire can fail to be realized. Similarly, a computer must be directed at the game in a certain sort of way for what appears to be a belief to truly be a belief. For instance,

merely moving a piece that, in a vacuum, constitutes an intelligent move does not mean that the computer was in a belief state. Rather, the computer must be directed at the game in such a way that the state it was in was about the game and that the belief had the potential to be wrong. To be clear though, merely having the potential to be wrong does not make the state Intentional. It is the way in which the state is about the world when it is at odds with the world that makes it an Intentional state. An electronic door failing to open when a person stands before it does not entail the door having a capacity for perception. It is both having states that can fail to capture something in the world, while being characteristically about the world, that would determine whether a chess "player" is playing. As such, the chess computer's agency may be settled by whether or not it has these sorts of states that are directed at the world with the feature of possibly being at odds with the world they are about.

VII. Case II: Language Speaking

Language acquisition is one of the most discussed abilities in the philosophy of Artificial Intelligence. This seems attributable to the Turing test, which was briefly reviewed in the introduction. The test aims to determine if a computer has a capacity for intelligent behavior by judging whether responses to questions are sufficiently human-like. Effectively, the design of the test is intended to determine whether a computer can speak human language. What is immediately striking about this test is its employment of something resembling Dennett's Intentional stance. That is to say, although the test does not *explicitly* take on this stance, by only analyzing the outputs of the computer and making judgments about those outputs, the test determines intelligent behavior by ascribing Intentionality to the machine. What this means is, if one were to ask a computer a question such as "what color is the keyboard I am typing on right

now?" and the computer were to respond to this question the statement, "the keyboard you are typing on is black" the only information that determines the computers language ability is *that statement*. This statement alone, and its felicitousness, establishes the language capacity of the computer. This is because the test, which is to say the evaluator, presumes a certain set of responses to the question being asked. In linguistics this is called the illocutionary effect.[11] If the computer outputs an expected response consistently to the questions asked of it, then the computer passes the test and appears to understand language.

But, the issue with the Turing test is the same as that with our clone chess player. What the evaluator is doing when reading statements the computer provides is ascribing Intentional states to the computer. When the computer outputs the statement, "the keyboard you are typing on is black," one could understand it as the computer communicating a belief state that it has. If a standard human perceiver in standard conditions were to perceive the keyboard in question, then she would perceive it as black. When the human perceiver is asked what color she believes the keyboard to be, she may tell the inquirer that she believes the keyboard to be black. However, what must be true of the human when she offers this statement is that she was indeed enjoying an Intentional state, namely a belief state, since under our framework she is necessarily related to the world in an Intentional way. These belief states, as well as states of desire and perception, seem to characterize the essence of agential language speaking. When humans speak languages, they do so in an Intentional way. For instance, a human speaker has beliefs about what her interlocutor is trying to convey by some utterance, has perceptions of certain inflections in the words she hears, or has desires to communicate particular concepts in her utterances. Therefore, for a computer to be agentially speaking, it must have Intentional states such as the ones delineated above.

[11] The effect that a certain utterance has on the hearer, and what is the expected response to the utterance.

As before, it must be made clear though that in deciding a computer's capacity for agential speaking, the question is not whether computers perform the task in the same way. Though technology has yet to reach the point where computers appear to be speaking a language to the same extent as a human, if technology were to develop to this extent it is likely that the way a computer parses out meaning of a sentence would be different than how a human does so. Likely, a computer would have to formulate syntax trees and reference databases that contain semantic meanings. Though it is possible humans speak in this manner,[12] but in such a way that formal procedures such as forming syntax trees only occurs in the background, it should not matter to us in deciding whether computers are mimicking or agentially doing the relevant action. Rather, all that should concern us is the relationship between the computer and the world when it appears to be speaking. We can bring this out further by the two thought experiments examined in this essay, Searle's Chinese Room and Davidson's swampman.

Turning to both the Chinese Room and Davidson's swampman, we can parse out in what way the computer should be related to the world if we are to judge it to be speaking language. Julia in the Chinese Room does not exhibit understanding because she has no understanding of the words she is appearing to translate. Though from the outside she appears to be speaking language, she is not about or directed at the concepts her output words convey. This failure to be directed at her words is similar to Davidson's swampman lacking an ability to be directed at the world Intentionally. Julia cannot form beliefs about the words she reads or form desires to convey X concept since she has no understanding of the characters she is appearing to translate. To further draw out the comparison, Davidson's swampman also does not have the capacity to be directed at "his" house or any other object he is associated with. As such, it becomes clear that

[12] I do not believe that human language speaking occurs in this way, and Dreyfus would likely agree that if humans and computers were to speak languages they would not do so in the same way.

Julia does not have the right relationship with the world in order for her to speak language. Since Julia does not have the right relationship with the world for her to speak language, then computers must not have the relationship with the world that Julia does if one is to judge them to be speaking language. They must be about the words in an Intentional manner, that is formulate belief, desire or perceptual states.

Furthermore, these states must be able to be at odds with the world. This is an inherent aspect of an Intentional state, and so should be apparent in the states the computer enjoys if it were to speak. That is to say, humans can have wrong beliefs about what their interlocutor intends to convey, as the hearer may believe the speaker meant one thing when really another thing was meant entirely. In such a case the hearer with such-and-such a belief state is at odds with the relevant state of affairs in the world. Similarly, when a speaker wishes to convey some concept in her head, this desire may sometimes fail to be realized as the hearer interprets the meaning of the utterance in a completely unintended way. As human speakers are Intentionally directed at the world when they speak, computers must be in a similar sort of relationship with the world in order to be judged to be doing the same action as a human. To be an agential speaker, it is not sufficient to consistently output the relevant response. Rather, the computer must have Intentional states when speaking that have the possibility of being at odds with the world.

VIII. Conclusion

The debate surrounding the issue of whether computers can perform human actions, as it currently exists, is at a veritable stalemate. Philosophers subscribing to each of the three camps considered in this essay all hold specific reasons explaining why or why not computers can be

agents in actions such as playing or speaking. The Machine Functionalists are primarily concerned with the causal nature of mental states. This feature of the human mind indicates that humans and computers function in the same way. The Biological Naturalists, however, do not respond to this point and instead advocate for the essential nature of a mind – that it must be biological. The picture becomes even more convoluted with Contextualism – positing the need for embodiment, which in turn allows for the faculty of skilled action. As the considerations from each of these three accounts causes an intolerable divide in the philosophy of Artificial Intelligence, some reconciliation must be made.

This essay has attempted to provide just that: a necessary condition that computers must satisfy in order for them to be judged to be agential actors in relation to relevant actions. For a computer to be doing such actions, it must have Intentional states, which is to say the computer must be related to the world in the same way a human player, speaker, and so on, is. The computer must be in a relationship with the world such that it is about the world in the same way a human agent would be when doing the same action. The computer should enjoy Intentional states such as beliefs and desires when playing chess or speaking language. However, though this essay makes an argument as to what the relevant consideration should be when judging a computer on its capacity for human action, i.e. Intentionality, it has omitted a method for concluding whether the computer has Intentional states. Discovering an approach to this issue I leave for future investigations, though I believe that progress can be made towards this by more closely examining how it is that Intentional states can be at odds in the world. In other words, I think that further consideration into how Intentional states "get it wrong", as it were, may demonstrate that there is something unique about the way in which Intentional states can be at odds with the world, versus a more ordinary sort of "getting it wrong".

IX. Bibliography

Brentano, Franz. *Psychology from an Empirical Standpoint*. Routledge, 2015.

Davidson, Donald. "Actions, reasons and causes." *Journal of Philosophy*, vol. 60, no. 23, 1963, pp. 685-700.

Davidson, Donald. "Knowing One's Own Mind." *Proceedings and Addresses of the American Philosophical Association*, vol. 60, no. 3, 1987, pp. 441–458.

Dennett, D. C. "Intentional Systems." *The Journal of Philosophy*, vol. 68, no. 4, 1971, pp. 87–106.

Dreyfus, Hubert L. *What Computers Can't Do: the Limits of Artificial Intelligence.* Harper & Row, 1984.

Fodor, A. Jerry. "The mind-body problem." *Scientific American*, vol. 244, 1981, pp. 114-124.

Hobbes, Thomas, and J C. A. Gaskin. *Leviathan*. Oxford: Oxford University Press, 1998.

Maloney, J. Christopher. "The Right Stuff." *Synthese*, vol. 70, no. 3, 1987, pp. 349–372.

Merleau-Ponty, Maurice. *Phenomenology of Perception.* London: Routledge, 1962.

Putnam, Hilary. "Minds and machines." *Journal of Symbolic Logic*, New York University Press, 1960, pp. 57-80.

Putnam, Hilary. "The nature of mental states" *Art, Mind, and Religion*, Pittsburgh University Press, 1967, pp. 223-231.

Searle, John R. *Intentionality: an Essay in the Philosophy of Mind*. Cambridge University Press, 2004.

Searle, John R. "Minds, Brains, and Programs." *The Behavioral and Brain Sciences*, 1980, pp. 417–457.