On The Possibility of an Artificial General Intelligence

By Jared Dembrun, September, 2016

Introduction

HAL, J.A.R.V.I.S., Data, The Matrix Architect. These characters are, in this day and age, mere figments of our imagination. They seem like such advanced machines that they can only exist in the distant future, but many modern experts believe that we will see at least primative versions of these machines within our lifetimes. Some even think these characters, or machines *like* these characters, will exist within that time frame. But what makes these experts so confident, especially when their predecessors of the 1950s got the timeline so embarrassingly wrong (Armstrong)?

In fact, the work done by AI scientists will undoubtedly lead to greater advancements in Artificial Intelligences, and these AI's may even begin to work in more generalized fashions than they do today, but a truly intelligent machine, a machine possessing what one could rightly call an “intellect,” cannot, in principle, exist. Such a machine must necessarily, as we will see, transcend materiality to some degree, at which point it could no longer be rightly called a machine, and its “intelligence” no longer rightly called “artificial.”

To illustrate this point, we will consider arguments from various philosophical perspectives as to the reason why such a machine cannot exist. In the first place, we must make a distinction between materialism, or the idea that everything which exists is made up of matter, which can loosely be defined as particles having mass, and dualism, or the idea that at least some parts of at least some things which exist are *not* made up of matter, or that they are immaterial. We will see compelling arguments against the possibility of an Artificial General Intelligence (AGI) from both materialist and dualist perspectives.

Mills, Machines, and Chinese Materialists

Gottfried Leibniz lived in the 17th century; he is most commonly known as the guy who was almost known for inventing Calculus (Mastin). In addition to his contributions to Mathematics, he was also a philosopher. In his famous work, *The Monadology*, Leibniz briefly discusses the possibility of a “thinking machine,” which, in his thought experiment, is “a machine, so constructed as to think, feel, and have perception, it might be conceived as increased in size, while keeping the same proportions, so that one might go into it as into a mill. That being so, we should, on examining its interior, find only parts which work one upon another, and never anything by which to explain a perception. Thus it is in a simple substance, and not in a compound or in a machine, that perception must be sought for.” (Cole, 2.1)

Leibniz's point may be difficult to see here, but he is simply contrasting the perceived behavior of the machine (ie, that it is thinking, feeling, and/or perceiving) with what it is *actually* doing, which is nothing besides moving parts around in accordance with physical laws. It is simply that the effects of this moving of parts make it seem that the machine thinks, feels, and perceives. Finally, Leibniz claims that there is nothing in this movement of mechanical parts which explains perception, and so concludes that it must be in a simple substance, rather than in complex substances, such as machines, that the explanation of perception is to be found. Of course, Leibniz, with his theory of Monads, was a dualist. Even so, his thought experiment, commonly titled “Leibniz's Mill,” is an important antecedent to the even more famous Chinese Room thought experiment by John Searle, which is an attempt to show that even from a materialist's perspective, AGI is impossible.

Another important antecedent to the Chinese Room was Turing's Paper Machine. The paper machine was really just a person who was given a set of instructions (eg how to play chess) written down on paper, who then followed the instructions. Based on this thought experiment, Alan Turing was confident, in the 1950s, that machines would soon be able to converse with human beings at the same level of intelligence that human beings converse with other human beings. Many of us are still waiting for these machines sixty years later.

Later, an AI researcher by the name of Roger Schank “developed a technique called “conceptual representation” that used “scripts” to represent conceptual relations (a form of Conceptual Role Semantics). Searle's argument was originally presented as a response to the claim that AI programs such as Schank's literally understand the sentences that they respond to.” (Cole, 2.2)

Perhaps the most important precursor to Searle's Chinese Room was Ned Block's Chinese Nation, an attempt to show that functionalist explanations of mental states are absurd.

> In “Troubles with Functionalism”, also published in 1978, Ned Block envisions the entire population of China implementing the functions of neurons in the brain. This scenario has subsequently been called “The Chinese Nation” or “The Chinese Gym”. We can suppose that every Chinese citizen would be given a call-list of phone numbers, and at a preset time on implementation day, designated “input” citizens would initiate the process by calling those on their call-list. When any citizen's phone rang, he or she would then phone those on his or her list, who would in turn contact yet others. No phone message need be exchanged; all that is required is the pattern of calling. The call-lists would be constructed in such a way that the patterns of calls implemented the same patterns of activation that occur between neurons in someone's brain when that person is in a mental state—pain, for example. The phone calls play the same functional role as neurons causing one another to fire. Block was primarily interested in qualia, and in particular, whether it is plausible to hold that the population of China might collectively be in pain, while no individual member of the population experienced any pain, but the thought experiment applies to any mental states and operations, including understanding language. (Cole, 2.3)

Finally, there is Searle's actual Chinese Room thought experiment, which asks us to imagine a situation in which there is a man in a windowless room, surrounded by various documents which instruct him how to change certain symbols and combinations of symbols into other symbols and combinations of symbols. He does not know what the symbols mean, nor must he even known that they are Chinese characters, much like in Turing's Paper Machine. He simply follows the instructions when prompted. This prompting comes from slips of paper which come in under the door. Once he has followed all of the rules to properly transform the symbols on the paper into new symbols, he slips the new paper back under the door. On the other side of the door is a Chinese speaking man, who asks questions of the room in fluent Chinese by slipping papers with the questions written on them beneath the door, and receives coherent and reasonable answers from the room also written in fluent Chinese. (Cole, 3)

Searle then asks us, where is the intelligence? Where is the understanding? Surely, the man does not understand Chinese, as we were told at the beginning. And it would certainly be silly to think of documents or even entire rooms as being “intelligent.” So then where is this supposed intelligence to be found? If one agrees with Searle that there is no intelligence to be found in this system, then he must agree that the same is true of any mechanical or electrical system which could conceivably be built, for all these systems do is the same as the man in the room. They receive input, follow instructions for transforming that input into an output, and output it. Of course, it will come as no surprise that there are many objections to the Chinese Room thought experiment. We will explore what seems to be the strongest objection.

The Systems Reply is the most common, and perhaps the strongest, reply to the Chinese Room argument. The Systems Reply is essentially that, while it is true that *the man* in the Chinese Room does not understand Chinese, the system *as a whole* indeed does. Proponents of the Systems Reply claim that this larger system, of which the man is only a component, specifically, the “implementer” of the algorithm, has the capacity to understand things which its individual constituents cannot. Searle rebuts the Systems reply by taking his thought experiment one step further. He asks us to imagine that the man in the room has memorized the steps, something he can, in principle, do. “He could then leave the room and wander outdoors, perhaps even conversing in Chinese. But he still would have no way to attach “any meaning to the formal symbols”. The man would now *be* the entire system, yet he still would not understand Chinese. For example, he would not know the meaning of the Chinese word for hamburger. He still cannot derive semantics from syntax.” (Cole, 4.1)

In response to Searle's rebuttal, John Haughland wrote in 2002 “that Searle's response to the Systems Reply is flawed: “…what he now asks is what it would be like if he, in his own mind, were consciously to implement the underlying formal structures and operations that the theory says are sufficient to implement another mind”. According to Haugeland, his failure to understand Chinese is irrelevant: he is just the implementer. The larger system implemented would understand—there is a level-of-description fallacy.” (Cole, 4.1)

It seems, however, that Haughland has not refuted Searle at all. He merely begs the question against him by asserting that the man in this case is just the implementer. Unlike the previous rebuttals, where the man could be taken as the implementer, and the entire room and all it contains as a whole system, in this case the man is all there is. The question is posed to the man, directly, in Chinese, and the response is likewise given by him in Chinese. No bit of information is output to some other part of any system outside of the man himself, and no additional information is received by the man in the process of responding. Yet, given what we know about him, we would call the man a liar if he claimed to know, or *understand*, Chinese. So Searle's analysis is correct; the man can *never* derive semantics from syntax.

So it would seem that Searle's argument holds up to scrutiny, or at least it remains undamaged by the argument which most critics use to attack it. But, there is a larger metaphysical debate involved in all of this which has yet to be discussed. Even if one believes that the truth of materialism implies necessarily the possibility of AGI, there is still the dualists' perspective to consider.

Two Dualists

In this section, we will examine arguments put forth by two dualists, Doctors James Ross and Edward Feser. These arguments, if they work, demonstrate that thought *requires* some immaterial substance, or aspect, the same kind of substance Leibniz believed was necessary for perception. And, if it is true that thought requires at least some immaterial aspect, then it can never be designed into a machine, which is necessarily completely material.

In his paper, titled, “Immaterial Aspects of Thought” from The Journal of Philosophy, Ross begins by speaking of the natural sciences and their task in assisting the philosophical field of epistemology in explaining, in terms of physical laws, animal cognition. He then goes on to talk about the larger and more daunting project, which he claims has hit a proverbial wall, of explaining *human* cognition in terms of physical laws, specifically the parts associated with mathematical reason and truth values. He writes

> That project seems to have hit a stone wall, a difficulty so grave that philosophers dismiss the underlying argument, or adopt a cavalier certainty that our judgments only simulate certain pure forms and never are real cases of, e.g., conjunction, modus ponens, adding, or genuine validity. The difficulty is that, in principle, such truth-carrying thoughts cannot be wholly physical (though they might have a physical medium), because they have features that no physical thing or process can have at all. (Ross, p. 136)

Ross goes on to show how formal patterns, like squaring a number, can be mimicked by purely physical machines for *particular instances* of those patterns (eg in the case of four raised to the third power), but that the definite pure form of such patterns cannot exist in these machines so long as the number of possible cases goes on to infinity, as it does in Arithmetic Operations, for example. To better understand this, let's see what Ross says about understanding addition:

> Adding-genuinely adding, not estimating-is a sum-giving thought form for any suitable array of numbers. If I add two "elevens," I am doing what would have given "forty-four" had I been adding two "twenty-twos" (and not making mistakes), and so on for every other combination of suitable numbers. I cannot be really adding when I do something which gives the "right output" but which cannot, by its form, determine the "right outcome" for any case whatever, even one on which I make a mistake. There is a

great difference between adding incorrectly and doing something else, like guessing, estimating, or following a routine or algorithm.

The adding I am talking about, like conjoining, is a form of understanding. This is not a claim about how many states we can be in. This is a claim about the ability exercised in a single case, the ability to think in a form that is sum-giving for every sum, a definite thought form distinct from every other. When a person has acquired such an ability is not always transparent from successful answers, and it can be exhibited even by mistakes. (Ross, p. 139-140)

And, Ross also argues that these forms (eg, addition, modus ponens, squaring, etc) *must* be "pure," "otherwise, they will fail to have the features we attribute to them and upon which the truth of certain judgments about validity, inconsistency, and truth depend" (Ross, p. 137). He then goes on to show that, although these functions are determinate, no physical functions or processes can be determinate. In support of this assertion, he offers the following argument.

Whatever the discriminable features of a physical process may be, there will always be a pair of incompatible predicates, each as empirically adequate as the other, to name a function the exhibited data or process "satisfies." That condition holds for any finite actual "outputs," no matter how many. That is a feature of physical process itself, of change. There is nothing about a physical process, or any repetitions of it, to block it from being a case of incompossible forms ("functions"), if it could be a case of any pure form at all. That is because the differentiating point, the point where the behavioral outputs diverge to manifest different functions, can lie beyond the actual, even if the actual should be infinite; e.g., it could lie in what the thing would have done, had things been otherwise in certain ways. For instance, if the function is x(*)y = (x + y, if y < 10^40 years, = x + y + 1, otherwise), the differentiating output would lie beyond the conjectured life of the universe. (Ross, p. 141)

What Ross is essentially saying here is that, given a possible infinity of inputs and outputs, no machine can offer a determinate function which will output the correct output given any input. We, however, understand these functions (such as *modus ponens*, addition, squaring, etc.), and *can* give the correct output given any input *just because* we understand the function itself.

We can see that Ross is correct in stating this at least about adding machines by examining the hardware involved in simulating this process. Take an electronic binary adder, for example. Such a device is a piece of electronic hardware which simulates the act of adding. But, we know that adders are constrained by their construction to two maximum inputs and one maximum output. For example, a four-bit adder, which takes in two binary numbers of at most four digits, and outputs one binary number of at most five digits, is constrained to the size of number which can fit in the given number of binary digits. Given any larger numbers, it will return an answer which is *not* a sum of those two numbers. Even if we were to somehow use all of the matter in the universe to build an x-bit adder, where x is the greatest number of possible wires we could build going in and out of the adder after first constructing such an adder and all the other components necessary to supply current to each wire, this adder would return the wrong answer given a number which requires x+1 binary digits in order to be written in binary as input. Yet, an average 10-year-old human being could add one to such a number with ease. This argument for the necessity of an immaterial aspect to thought is quite strong.

Dr. Feser offers a different argument in his book *The Last Superstition: A Refutation of the New Atheism*, which is a refutation of the typical materialist's account of the mind. He begins by describing the basic picture of the mind given by a materialist:

Now, let's consider the dominant materialist approach to explaining the mind in purely "naturalistic" terms, according to which the brain is a kind of digital computer and the mind is the "software" or "program" that is implemented on this "computer." … Individual thoughts are just physical symbols in the brain – like words or sentences, only encoded in the form of neural firing patterns, rather than in ink (as when you write a word or sentence), or sound waves (as when you speak it), or magnetic pattern on tape (as when you utter it into a tape recorder), or electrical current (as when you type it into a computer).

> Thinking – going from one though to another – is just transitioning from one symbol in the brain to another according to the rules of an algorithm, in just the way a pocket calculator goes from "2" and "+" and "2" and "=" to "4" according to the rules of an algorithm, the difference between the calculator and the brain being a difference in degree but not in kind. The symbols get their meaning from the cause-and-effect relationships they bear to objects and events in the world outside the brain; hence such-and-such a brain process will count as a symbol meaning "There's a snake!" if it was caused by snakes affecting the sense organs of the speaker in such-and-such a way, and/or because it was hardwired into the brain by natural selection, since it got people to avoid snakes and this behavior was conducive to their survival. (Feser, 238-239)

Next, Dr. Feser describes some of the absurdities and inconsistencies in such a position:

> Here are some of the absurdities. First of all, nothing counts as a "symbol" apart from some mind or group of minds which interprets and uses it as a symbol. For example, the words you're reading right now count as words at all only because English-language users so count them, given a series of historical accidents as a result of which "cat" is conventionally used to refer to cats, "dog" to dogs, and so on and so forth. In themselves and apart from these conventions, "dog," "cat," and the like are just meaningless squiggles of ink or meaningless noises. The same thing is true of every other physical symbol… But then it is true also of any "symbols" purportedly encoded in the brain: By themselves they cannot fail to be nothing more than meaningless neural firing patterns (or whatever) until some mind *interprets* them as symbols standing for such-and-such objects or events. But obviously, until very recently it never so much as occurred to anyone to interpret brain events as symbols, even though (of course) we have been able to think for as long as human beings have existed. It follows that no brain events could have *been* symbols of any sort for all this time, in which case our thought processes were not the mere processing of symbols in the brain. More to the point, since the materialist's "computer model" of the mind tries to explain the mind in terms of symbols in the brain, but nothing counts as a symbol in the first place except when interpreted as such by a mind, the theory goes around in a circle and is simply incoherent. (Feser, p. 239-240)

Feser accuses the materialist of reasoning in a circle when he asserts that the mind can be explained in modern computing terms. This is because symbols themselves rely on minds, but the mind (taken as a computer) is supposed just to operate symbolically. And, indeed, modern computers cannot operate in any way *other than* symbolically. Even the "machine learning" algorithms used to produce systems like Watson, which can accurately diagnose medical patients, are reducible to symbols, since the hardware on which the software runs is purely symbolic. But these symbols, both the binary values in the CPU register, RAM, and persistent memory device(s), and the English text or speech output by the machines, have meaning *only because* our minds give it to them. Otherwise, they would have no more meaning in them than a particular arrangement of seashells washed up on a beach by the tide, as both are simply processes beholden to mechanistic laws. So it would seem that the case for an immaterial aspect to thought, and thus the impossibility of AGI, is doubly strong.

A Poetic Death to AGI

And now, if the reader is not yet thoroughly convinced of the impossibility of AGI, I would put forth my own argument, neither as strong nor as bold as Searle's, Ross's, or Feser's, yet still capable of casting considerable doubt upon the possibility of constructing an AGI. Consider the following poem by Thomas Morley:

April is in my mistress' face,
And July in her eyes hath place;
Within her bosom is September,
But in her heart a cold December.
(Morley)

What does this poem *mean*? Is it meant to state that the narrator's mistress has these months literally *in* her various body parts? Of course not; we know that it is a poem, and that of course the literal interpretation of

the words is the incorrect interpretation. But let us imagine the steps that an AGI would have to take just to determine that the words are not meant literally in the first place, and then let us see how such a machine might attempt to interpret the poem.

First, how would an AGI, in theory, determine that this is indeed a poem with some figurative language? Well, it would need enough background knowledge to know what poems are, and to know that some words in poems are meant literally, and others are meant figuratively. This does not seem like a whole lot of information, but remember, this machine needs to be able to do this *for each task humans can do*, of which interpreting poetry is only one. Still, it seems that it would be, at least in principle, possible to design a machine with enough background knowledge that it could figure out that this particular piece of literary work is a poem. Indeed, it is likely that modern machine learning algorithms could be made to differentiate different kinds of literary works, poems being one of them.

Now, our hypothetical machine may know that this is a poem, and it may know that at least some words in some poems are meant figuratively, while others are meant literally. But, how can it determine *which* words in particular are meant in which way? Well, we can suppose that our machine has *even more* background knowledge. It would need knowledge *at least* about every noun it encounters, as to whether those nouns are often used in poetry figuratively or literally, whether they make sense in their places here literally, and also all of the different ways they could be used figuratively, and whether any of *those* ways makes *more* sense in each case than the literal interpretation. Designing such a machine seems to be quite a daunting task, and would likely be a never-ending process as language evolves over time. Still, it is, in principle, possible, though it is unlikely that mankind could ever design such a machine and have it work in a way that we would consider “intelligent.”

But let us assume for the sake of argument that just such a machine has been built, and it can point out each noun in a poem and accurately determine which are meant figuratively and which are meant literally. Now, it must somehow interpret the overall meaning of the poem. But take the first line: “April is in my mistress' face.” The machine has somehow determined that the word “April” here is neither being used to refer to a woman named April, nor to the month of the year, but that it is somehow describing a feature of this mistress' face. Never mind the fact that it also has somehow determined that the words “mistress'” and “face” are being used as they typically are.

With which word or words would the machine replace the word “April” to make sense of this poem? Would it use the word “beauty”? If so, then how does the machine distinguish between the *concept* of beauty being literally in the mistress' face, as opposed to the mistress being beautiful? Now it must make a decision about which nouns are only concepts, and which are concepts which refer to actual, physical objects. Or say it used a different word, like warmth or life. How does it distinguish between the mistress' face containing warmth (ie has a substantial amount of heat) or life (being alive) and “warmth” or “life” taken to be positive and attractive qualities when describing a person's physical features?

Both the difficulty in determining which words are meant figuratively and how to interpret those figurative meanings could be considered a kind of filter, not unlike the proposed “Great Filter” invoked when some philosophers attempt to explain the Fermi Paradox. Both of these challenges, while not in principle insurmountable, are incredibly daunting, and their existence should cast considerable doubt on the likelihood of mankind ever constructing such machines. These challenges apply not only to poetry, but to everyday language, where people are often figurative, sarcastic, and implicit in their communication, and the ability to properly distinguish implicit meaning in words and sentences would be of the utmost importance to any AGI. Indeed, it would be a *necessary* component.

Conclusion

It would seem that the case against AGI has been thoroughly made, first from a materialist view, with Searle's Chinese Room argument, and also from a dualist's view, in which both materialism itself and the possibility of AGI *de facto* are refuted. Further, even if we (against the sound conclusions of the above arguments) take the building of AGI to be something which is, in principle, *possible*, we still have very good reason to think that the way we interpret language would pose an insurmountable, or nearly-insurmountable obstacle to the actual *development* of an AGI. All things considered, the future of AI's as super-intelligent overlords or benevolent dictators to the human race, or even crew members and/or officers aboard futuristic space vessels, seems incredibly fantastical, and altogether very unlikely, if not downright impossible. Still, the possibilities in the field of Artificial Intelligence are vast, and the knowledge gained in the field of computer science more generally by attempting to mimic the minds of humans in machines need not be discounted. It is simply that we will never have an Artificial Intelligence capable of thought on the same level as human beings.

Bibliography

Armstrong, Stuart. “AI Prediction Case Study 1: The Original Dartmouth Conference” *LessWrong*. March 2013. http://lesswrong.com/lw/gue/ai_prediction_case_study_1_the_original_dartmouth/ Accessed: September 2016

Ross, James. “Immaterial Aspects of Thought” *The Journal of Philosophy.* Vol 89, No 3. March 1992. (Note: free copy hosted by University of Notre Dame at http://www3.nd.edu/~afreddos/courses/43151/ross-immateriality.pdf)

Cole, David. “The Chinese Room Argument” *Stanford Encyclopedia of Philosophy*. March 2004 (Rev April 2014). http://plato.stanford.edu/entries/chinese-room/

Mastin, Luke “17th Century Mathematics – Leibniz” *The Story of Mathematics.* 2010 http://www.storyofmathematics.com/17th_leibniz.html Accessed: September 2016

Feser, Edward. *The Last Superstition: A Refutation of the New Atheism.* St. Augustine's Press, United States of America 2008.

Morley, Thomas. “April is in My Mistress' Face” pub. 1594.