Could Machines Think ?

Brito Francisco (PH), Kapoor Sayash (IC), Mucchietto Andrea (STI), Novello Alessandro (STI)

Project SHS 1st year master

Supervised by

Esfeld Michael-Andreas, Philosophy of Science
Mario Hubert, Philosophy of Science

Report accepted on [20.12.2017]

Lausanne, academic year 2017 – 2018

word count: 7015

COULD MACHINES THINK?

BRITO FRANCISCO, KAPOOR SAYASH, MUCCHIETTO ANDREA, NOVELLO ALESSANDRO

DECEMBER 18, 2017

Abstract. The dawn of AI has made it tempting to think of machines as being conscious, intentional, *thinking* entities. Starting from the concept of the Turing test, we analyze John Searle's Chinese room *gedankenexperiment*, and Roger Penrose's conjecture about the importance of as-yet-unknown physics in the description of a conscious brain. We review John McCarthy's and Aaron Sloman's arguments in favor of the so called strong AI thesis, and close with some general remarks about all the aforementioned work, and some of our own thoughts on the subject.

Contents

Part 1. The rise of AI - question in context

Could a machine think? Before we tackle the question, let us first make some fundamental remarks. Objectivity and subjectivity are systematically ambiguous regarding the distinction between an epistemic sense - relating to knowledge - and an ontological sense - relating to existence. The point is that there are both objective and subjective statements in both an epistemic and an ontological sense.

In fact the ontologically subjective character of consciousness does not in principle impede it from being described by an epistemically objective scientific theory, although there is currently no widely accepted framework to describe it. Moreover there is a crucial distinction between observer independent and observer relative phenomena. Human consciousness belongs to the latter, containing an element of ontological subjectivity. However, as we have already remarked, an epistemically objective scientific theory can describe an observer relative domain. Consciousness is in principle prone to a scientific description.

Phenomena that are studied in cognitive science - intelligence, cognition, memory, thought, or perception - can have two different senses: an observer independent and an observer relative sense. The discussion is crucial because observer relative phenomena serve as reference for many important concepts in cognitive science.

These ideas serve as a bedrock to John Searle's (Searle 1980) attack on the hypothesis that a machine could think - strong AI, a view that is mostly shared by Roger Penrose (Penrose 1989). Intentionality is a product of causally connected brain processes. Therefore a computer program instantiation cannot be a sufficient condition for intentionality. The only hope for what is called strong AI is to attempt to replicate the causal powers of the brain. Searle's famous Chinese room argument was constructed in response to what was at the time a widely spread and oversimplified belief that mental states could soon be reproduced in computers.

The Chinese room argument appears more than thirty years after Alan Turing's hopeful defense of the future possibility of ascribing consciousness to a machine. Attempting at establishing criteria for this, Turing set up the *Turing test* thought experiment. The reasoning is described in a seminal paper which we shall review(Turing 1950).

Science and philosophy go hand in hand, and a purely operationalist view can hinder the galloping progress of the ever growing fields of computer science and artificial intelligence. What were previously deemed as hypothetical concerns of a distant future could quickly become the ethical dilemmas of our own time.

The question is related to what is called philosophy of the mind. A good introduction to the topic is Searle's book (Searle 2004), from where many fundamental ideas in this essay were drawn from.

Part 2. Turing Machine

1. The Imitation Game

In his seminal paper *Computing Machinery and Intelligence*, Alan Turing starts with the question 'Can machines think?', but finds this approach unreasonable to find an insight into machine intelligence. He instead proposes an imitation game to be played by three entities – a machine (A), a human (B), and an interrogator (C). The interrogator tries to determine which of A and B is a machine by asking questions to A and B. Turing then asks the question: "Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?" (i.e when A is a woman and C has to determine which of A and B is a man). This question replaces the original question 'Can machines think?' According to Turing, it is worth changing the original question so that we can express it unambiguously. The question 'Can Machines think?' requires one to define 'machine' and 'think', however, the meaning of these terms is subjective, and finding a general consensus would be both difficult, and time-wasting. Thus, Turing proposes the removing of this ambiguity by changing the question we are trying to answer, from 'Can Machines think?' to 'How well does a machine perform in the Imitation Game?'.

Machines in the game. Turing allows only 'digital computers' to take part in the game. He places this restriction to disallow arbitrary methods, for example 'cloning' a human and calling it a 'machine'. Further, he allows computers with 'infinite store', that is, computers having as large a storage as required, and changes the question to: "Let us fix our attention on one particular digital computer C. Is it true that by modifying this computer to have an adequate storage, suitably increasing its speed of action, and providing it with an appropriate program, C can be made to play satisfactorily the part of A in the imitation game, the part of B being taken by a man?"

2. Objections and Turing's counterarguments

The Theological Objection.

Objection. Thinking is a function of the soul, and God has given a soul to humans, but not to other animals or machines. Hence, no animal or machine can think.

Counterargument. While Turing considers this argument unacceptable, he nevertheless proceeds to answer this argument in the theological framework itself. He argues that the theological objection is restricting the omnipotence of the almighty, and that if 'He' (the Almighty) wants to confer a 'soul' to an elephant (or, by extension, a machine), 'He' could do so.

The "Heads in the Sand" Objection.

Objection. One refuses to acknowledge the possibility of a thinking machine, because the consequences would be too dreadful.

Counterargument. Turing doesn't deem the objection worthy of a refutation, since it isn't based on any foundation except Man's superiority complex.

The Mathematical Objection.

Objection. This objection applies to fundamental limitations of discrete state machines, such as Gödel's theorem, among others.

Counterargument. Turing counters this argument saying that while these are established limitations of machines, there is no result proving that similar limitations don't apply to humans. Further, he states that while there might be humans able to stump a *given* machine, the same cannot be said about *all* machines, that is, a human cannot be cleverer than *all* machines at once. Finally, Turing concludes the counterargument by saying that those who have this objection wouldn't mind the imitation game as a proxy for machine intelligence, unlike the believers of the previous two arguments.

Arguments from Various Disabilities.

Objection. This objection is aimed towards a specific aspect that a machine is said to lack - falling in love, having a sense of humor, doing something really new etc.

Counterargument. Turing argues that there is no support offered for these arguments, and people believe in these arguments for the sole reason that they have never themselves seen a machine performing these tasks. He further goes on to give specific counters to each of the examples mentioned, and feels that given enough storage and the ability to manage it, machines can do any of these.

Lady Lovelace's Objection.

Objection. Lady Lovelace stated in her memoir that Babbage's analytical engine couldn't originate anything, but could only 'do whatever we know how to order it to perform'.

Counterargument. Turing gives the following counters to this argument: first, he states that while it did not seem that computers existing in the times when the memoir was written (1842) didn't have that property, there was no certain proof that it *couldn't* happen some time in the future. Then, he goes on to question whether there is anything *new* under the sun at all, and whether things that we are surprised are because of our own pre-existing idea of what we should expect. Finally, Turing states a fundamental misconception which creates the view that machines cannot give rise to *surprises*. It is the belief that every logical consequence of a known fact is obvious upon knowing the fact, which leads to the false assumption that working out the consequences of given data is trivial.

Argument from Continuity in the Nervous System.

Objection. The nervous system is not a discrete-state machine, and thus we cannot expect a discrete-state machine to mimic it.

Counterargument. Turing responds to this argument saying that the interrogator will not be able to take advantage of this fact. He exemplefies this point by arguing that a digital computer could mimic another continuous device – a *differential analyser*, by perhaps adding an element of randomness in the output that the computer chooses.

The Argument from Informality of Behaviour.

Objection. The objection states that since there is no 'set of rules' to describe what a human might do in *every possible circumstance*, this implies that humans aren't machines. Turing reconstructs this argument as: 'if each man had a definite set of rules of conduct by which he regulated his life he would be no better than a machine. But there are no such rules, so men cannot be machines'

Counterargument. Turing counters this objection by arguing that we cannot make a conclusion about the absence of such laws in humans by empirical evidence alone. In return, he challenges anyone to find the underlying rule of a machine that outputs a 16 digit number, on being given a 16 digit number as input (which he argues is nigh impossible). Thus, we cannot say for sure that there are *no such laws* governing humans.

The Argument from Extrasensory Perception.

Objection. Humans possess Extrasensory perception (ESP), in the form of telepathy, clairvoyance, precognition and psychokinesis. This would allow the interrogator to distinguish between the human and machine, by perhaps featuring a human who is a good telepathic receiver as one of the witnesses, and observing which of the witnesses performs better in taking a guess about something known only to the interrogator.

Counterargument. Turing counters this argument by saying that if the digital computer in this case uses a random number generator, then it might also be able to give the correct answer, due to the psychokinesis of the interrogator. On the other hand, the human witness might simply tell the right answer using clairvoyance.

3. Learning Machines and the Human Mind

In the last section of the paper, Turing talks about how the Imitation game could be played successfully by a machine, and what would need to go into the creation of such a machine. He argues that to imitate the human mind, the machine can go through the same steps a child does when born – to develop an initial model of the machine, and then subject it to an appropriate 'course of education'. Further, he also proposes that a system of rewards and punishments, similar to how it appears in human education, could help expedite the process.

Part 3. Consciousness and programs

In his both highly praised and criticized article (Searle 1980), Searle aims to construct a *gedankenexperiment* compelling the reader to draw a subtle distinction. Even though the outcome of the interaction with an entity - whether it is a computer or not is unknown - may be human-like, that does not suffice to attribute intentionality to it. Searle stresses that cognitive states and mental processes can only be attained by a machine equivalent to a brain, that is having its causal powers. The existence of such causal powers cannot be evaluated solely on operationalist grounds. Indeed causality in the underlying process leading to the entity's response is the key to decide whether consciousness is to be ascribed to it or not. Instantiating a program is not a sufficient condition for intentionality. The mechanism of intentionality requires causal powers equivalent to those of the brain. Therefore building artificial intelligence must consist of more than devising a machine instantiating a properly designed program.

The discussion is centered around the Schank program. Aiming at simulating the human ability to understand stories (page 418), the program provides an answer to a question related to the story that is asked after telling it. This question should be about something that is never explicitly told in the story, but that comes as an obvious conclusion.

A furious man leaves a restaurant after being served a burned hamburger. In the same way that a human can infer that the man did not eat the hamburger, a Schank program can output a binary decision answering the corresponding yes or no question. All it takes is to define a representation so that the data printed by the machine is meaningful to the reader. The machine defined in this manner will then print out human-like answers to questions of this sort.

Searle proceeds to remark that the claim of strong AI is that a machine defined as such *understands* the story, effectively explaining the human ability to decide about answers to questions posed upon listening to and further understanding a story. He does not agree with this claim. Here it is crucial to stress that human understanding arises from a set of causally connected cognitive states, which he does not find plausible to attribute to the machine, providing strong arguments. However, as we will see, Searle's arguments can too be criticized, although he provides a strong benchmark or starting point to the modern discussion about the powers of AI.

4. The Chinese room

Let us reconstruct the thought experiment proposed by Searle. Its goal is to point out the absurdity in deducing that the machine understands simply by instantiating a Schank program.

Suppose you are not a Chinese speaker and are locked alone in a room. You will now instantiate a Schank program. You are provided with a page of Chinese writing, but you are not able to speak or write Chinese, and in fact you are not even sure that the document provided to you is a page full of Chinese characters. For all you know it could be some other language, or, as Searle puts it: "so many meaningless squiggles"(page 420).

Further suppose you are again provided with a similar page. In contrast, this time it comes accompanied by rules you can simply follow to correlate the current and the previous batch. The rules are in your native language and you understand them. Despite this elusive task, we mustn't forget that all you do is manipulate a series of formal symbols you recognize by their shapes. No meaning is attributed to any of these by you, while effectively instantiating the Schank program.

A third batch arrives, and with it a set of rules correlating it to the second batch. Now, besides that, you are given another set of rules allowing you to output Chinese symbols in response to this last batch.

It follows from simple association that it was not *understanding* by your part that led you to formulate what the Chinese people outside interpreted as answers. Call the first batch a script, the second a story and the third questions. Furthermore, call the set of rules to manipulate the formal symbols a program. Nothing of what you did was any different from what could have been performed by a computer.

The bilingual Chinese people on the other side also speak your native language, say portuguese, and they now give you exactly the same batches written in portuguese. In the hope that they will be able to distinguish you from a computer, you answer them. The program was so good though, that your answers are indistinguishable from those of the computer. Both are equally plausible.

It now becomes evident that while in both cases the outcome was the same, the inner process leading to it was hugely different. In the first, you simply manipulated without any interpretation the *formal* symbols that were given to you! This was done purely by performing computational operations on a specified set. In contrast, in the second you produced a sequence of cognitive states that led to an interpretation followed by the realization of an intention, that of replying.

Let us now examine the claims of strong AI in light of this thought experiment. In spite of the indistinguishability of your inputs and outputs from those of a Chinese speaker, like the computer instantiating the Schank program, you understood nothing when told the story in Chinese. It logically follows that a program cannot by itself explain human understanding. What is not so trivial is whether or not it constitutes a necessary condition contributing to human understanding, and ultimately awareness and consciousness.

Searle hasn't disproven the claim that formal symbol manipulation is involved in the case in which you understand a story when people send you the text in your native language. Such claim is merely deemed implausible as nothing in the formal symbol manipulation itself seems to be what has led to understanding. In fact they seem to be completely uncorrelated at least in this thought experiment. There seems to be

no compelling evidence that computational operations performed on formally defined elements are at the basis of consciousness. Is there any evidence at all that you are operating a formal program when you have a conversation in your native language, or, as a matter of fact when you carry out any mental process?

5. Causality and Intentionality

What do you have in the case where your native language is spoken that you do not have in the Chinese room? In what does understanding consist and why can't we give it to a machine? We begin by remarking that while one can understand to many extents, there is always a clear distinction between a complete lack and the existence, however limited, of understanding. One can speak many languages without being able to read the great epic poets of each. A machine is not even in the same class. It understands nothing, in fact it is simply formally manipulating elements of a specified set. The confusion is due to the fact that we naturally attribute our own intentions to objects in a metaphorical sense. If the computer instantiating a Schank program is said to understand in this metaphorical sense, there is no question to discuss. This is clearly not the claim at stake. The claim is that cognitive states in humans and computers can be made equivalent. As per Searle we would argue that no partial or incomplete understanding can be attributed to a computer; it straightforwardly does not understand.

A human brain has a specified chemical and physical structure that, under certain particular conditions produces "perception, action, understanding, learning, and other *intentional* phenomena" (page 427). The point is that to produce these phenomena we need to devise a structure with exactly the same causal powers of the mind that lead to intentionality. In principle, a system with different biochemistry or possessing as-yet-unknown physical properties could have these, but that is beside the point. Purely formal properties are not sufficient to constitute intentionality.

In our experiment we encountered causal properties that are specific of a particular realization of a formal model - the set of rules. These properties would differ or even not exist at all in a different realization of the model. Instantiate Schank's program on any device, be it a computer or a non Chinese speaking person. It becomes clearer that nothing in particular is *understood* about the Chinese message when you note that understanding seems to be dependent on the particular realization of the model. The entity instantiating the program does not and cannot understand. This simply follows from its definition.

One more misleading remark is worth being discussed. Surely, in the case of a real brain, one could formalize the sequence of synapses in a way that is similar to a computer. However, it is the actual properties of the sequences and the way that they are architectured that gives rise to mental states. In this sense, the strong version of AI is clearly mistakenly claiming that a formalization - that cannot by construction explain understanding - leads to understanding.

6. Objections and Searle's counterarguments

Systems reply. It is not the individual, but the system comprising the entity that gives out answers in Chinese that understands. That is why we are able to devise such compelling arguments about the individual not understanding: in fact by himself he does not.

It is quite straightforward to see that the subsystems constituting the large system can all be collapsed onto a single system, and once the division is broken the counterargument too breaks. For instance, imagine that the human has very good memory and memorizes all the instructions. Suppose we do in fact consider a subsystem of the man that understands Chinese and another that understands his native language. Although both pass the Turing test, understanding does not follow solely on the basis of this criterion.

Robot Reply. Thinking is not a matter of instantiating programs alone. A set of causal relation with the outside world may be added: in addition to containing Schank's program, the computer is put inside a somewhat anthropomorphic robot form that uses other programs to mimic human actions.

Clearly nothing is added in terms of an increase in understanding. The original lack of intentionality persists. For example, the robot may have a video camera enabling it to see and then act accordingly by a set of limbs controlled by its "brain". Suppose that transposing this again to the Chinese room experiment, some of the symbols represent the signal of the camera and others instructions to act accordingly. The robot is still only manipulating formal symbols and that continues to be insufficient to explain cognitive states.

Brain Simulator Reply. Suppose you design a program simulating the neuron firings of a Chinese speaker while understanding the message and giving answers. The program even includes parallel computing to more accurately account for the way that the brain presumably processes natural language (page 420). Can a machine instantiating this program be said to think?

We begin by noting that the claim of strong AI we are discussing is functionalist in nature: AI should require no knowledge of the behavior of the brain to explain the behavior of the mind.

Now replace the monolingual symbol shuffling man in the Chinese room by a man operating a system of water pipes and valves. Establish a correspondence between each water connection and a synapse in the Chinese brain. The system is constructed so that an answer appears at the end of the series of pipes. This answer is simply generated by turning on and off all the right valves.

No understanding was introduced. Neither the man nor the pipes understand. The conjunction of the two

again doesn't either. The man could in principle operate all this formal structure by imagining all the simulated water pipe neuron firings, collapsing the system again into a single element.

The simulator is simulating only one insufficient aspect of the brain in describing consciousness: its formal structure. The causal properties that are able to produce intentional states are neurobiological in nature and in this example they were clearly dissociated from the formal part, yielding no understanding.

Combination reply. Adding the previous three replies, we end up with a decisively convincing human-shaped robot behaving like such, that simulates our synapses correctly.

Just as we find it natural to ascribe intentionality to animals due to them being made of the same elements as we are - eyes, arms, legs, ... - we make the assumption that if a robot acts and looks like us, it must have a similar mechanism to our own, producing the mental states that allow interaction.

Ascription of such mental states breaks as soon as we are able to account for the behavior of the robot without assuming it to possess a similar mechanism to our own. This is particularly true if the mechanism turns out to be manipulating formal symbols.

Crucially, note that the hypothesis of intentionality was abandoned after realizing that the properties of the actual physical substance comprising the robot are irrelevant to its "understanding" ability. The program might just as well be realized in the water pipe system.

Many mansions reply. Analog and digital computers were the state of the art technology at the time. What if we eventually build the devices having the causal powers Searle claims to be at the heart of intentionality? Maybe in the meantime we did already. The argument seems to be specifically directed towards a particular state of affairs, and not to the general possibility of a more advanced AI producing intelligence and cognition.

No objections are presented to the question that is raised here. The problem is that it redefines AI relative to what was the starting point of this whole argument. In fact, AI is redefined as the abstract as-yet-unknown scientific framework that explains cognition. This is too loose a definition. With it, AI becomes a distant goal in the sense that attaining such a framework is no easy feat. The redefinition aims at deceiving one into thinking that the mission is simpler than what it actually consistently proves to be.

The argument presented here aims to challenge the idea that mental processes are symbolic formal operations. It no longer applies to challenge a different redefined thesis.

Part 4. Computers and the laws of physics

Roger Penrose argues that conscious thought processes may be explained in terms of physical processes of unconventional nature (Penrose 1989). We focus in an article published on *Behavorial and Brain Sciences* (Penrose 1990) as a follow-up to *The Emperor's New Mind*.

7. Some preliminaries

Penrose conjectures that although mathematical non algorithmic processes exist, no one has ever explored the possibility that they play a crucial role in the domain between classical and quantum phenomena. In particular, although currently there isn't a widely accepted theory of consciousness, there is a good chance that conscious processes are non algorithmic. Therefore, non algorithmic physical processes could be occurring within the brain. Furthermore, this implies that some aspects of the brain are not prone to simulation, in the sense of devising a program in the currently known framework.

The physicist has a peculiar view in that he too does not agree with strong AI, but he agrees with John Searle only to some extent. Searle allows for the possibility to simulate the brain, even though he stresses that that does not imply that the program running the simulation is conscious. This possibility rests on the fact that the physical processes within the conscious brain follow well defined mathematical operations. Penrose doubts the very computability of these operations, conjecturing that they might be of non algorithmic nature.

We start with the assumption that brains are governed by the same physical laws as everything else. This is justified by the fact that currently known minds are a product of the activity of brains, which are part of the physical world. The key point is that we believe to understand matter *well enough* to describe the brain, but Penrose believes it might not be the case. New physics reaching fundamentally new understanding may need to be developed before we hope to construct a working theory of the mind.

In the following sections we will mention some concepts of quantum physics. In particular, let us stress the intrinsically probabilistic nature of quantum mechanics, implying that a system is always in a *superposition* of quantum states. There is no classical physics, or common sense analogy for this. Things at small scales seem to behave according to strange rules. One of those is that measurements, or observations, affect the state of the system. When you measure a quantity of one of these systems that is in a superposition of states, it collapses into one of the states, according to a prescribed list of probabilities. This list can be computed using the first principles of quantum mechanics. The process in which the system "chooses" one of its possible states is called collapse of the wave function (which somehow contains the list of probabilities). Another fundamental concept is that physical quantities are discrete. This means that they come in packages, taking only on values separated by given fixed intervals. This affords an elementary particle description, in which the interactions between particles themselves are represented by other mediating particles that carry the "force", or more precisely, a message containing the

interaction, and its properties. There is an experimentally tested particle description for every fundamental interaction in nature, apart from gravity. The mediating particle carrying the information about gravitational interaction - the gravitational field - is called the graviton.

8. Real brains and model brains

Although it is not clear in which part of the brain consciousness resides, some areas really do seem to be purely automatic, like the cerebellum. Ironically, the cerebellum has a much higher density of neurons than the cerebral cortex, where processes are often of conscious nature. This again suggest that strong AI does not rest on the correct principle in the first place. A computer model of neuron firing may not be a good enough model to capture consciousness. Brain plasticity, for instance, the strengthening and weakening of connections between neurons is thought to be at least partially responsible for permanent memory. The neuron firing model could not ever capture this phenomenon.

Parallel computing appears to be a good candidate in brain modelling. Penrose believes that it still couldn't capture consciousness. He gives the intuition that the apparently omnipresent nature of consciousness is at odds with the principles of operation of computers, either serial or parallel. There is in fact no difference between what a serial and a parallel computer *can* compute.

The main point is that quantum superposition may be ingrained into the brain, now regarded as possibly well described by a quantum computer. However, the non computability hurdle remains with quantum computers. In spite of the relevance of many quantum phenomena in the workings of brains, no quantum description of the action of the human brain exists to this day. Even just by taking a closer look at the principles of quantum mechanics this seems unreasonable. This is because the computer would have to be "observing itself" continuously (page 653). This would entail continuous collapses of the wave function, and the current quantum theory does not account for this.

9. Physics of the mind

Consciousness appears to be an evolutionary feature. However, some brain processes are intentional, that is conscious; others are automatic, unconscious. The distinction between them could be that conscious processes are non algorithmic. Consciousness could then be described by precise but non computable physical laws. Suggestively, conscious processes seem to possess some globality about them. This may deem the search for an optimized "brain algorithm" hopeless.

What is proposed is to relax the algorithmic condition of strong AI. The conjecture is built upon brain plasticity. There are regions of the brain called dendritic spines, where particular connections between neurons exist. The size of these regions affects the strengths of these connections in very short time frames. At the most fundamental level this process is quantum in nature. Now, we may consider the combinations of strengths of the connections to constitute a quantum superposition. This would mean that the brain makes many calculations simultaneously, not in the sense of parallel computing, but in the intrinsically probabilistic sense of quantum mechanics. A conscious thought emerges when the wave function collapses. Such process is thought to be connected with quantum gravity. In particular the graviton, the particle quantizing the gravitational field seems to establish a measure of when the collapse of the wave function occurs. If shown to be true, this conjecture would mean that the emergence of consciousness would be necessarily of non algorithmic nature!

Part 5. Critical views in favor of AI

10. McCarthy

McCarthy's core argument in defense of AI is based on the *commonsense informatic situation*(McCarthy 1990). This relates to the methods that provide computers with human decision ability in daily life cases. The challenge lies in programming such situations within an already formalized theory. Issues arise from the following facts:

- Knowledge on general phenomena and on particular situations needs combining. General phenomena are ruled by the laws of physics, whereas particular situations are related to human abilities and human features.
- It is not known in advance which phenomena have to be taken into account.
- For problem-solving situations whose theoretical model is fully determined, computational complexity can force approximating the problem by systems whose laws are partially unknown.

Achieving human-level commonsense knowledge remains a first concern for AI theory. According to McCarthy, a path towards AI progress would involve:

- Representing more kinds of details and phenomena of the real world by logical formulas;
- Examining intellectual mechanisms;
- Representing the approximate thinking of human beings that is used in commonsense reasoning;
- Devising better performing algorithms to search the space of possibilities.

McCarthy focuses on *formalizing non monotonic reasoning*. Human reasoning is non monotonic. Consequently, a mathematical model whose set of conclusions does not monotonically increase with the set of premises is required (McCarthy 1990).

Recall the Chinese room argument. A man who does not know Chinese is assembling Chinese sentences following a book of instructions and rules. The program is embodied by the book and the man is the hardware interpreter. Hence the program should know Chinese in

order to engage into a nontrivial Chinese conversation. If the man had memorized the rules, then he would be interpreting another personality besides his own. This additional personality would be the Chinese one taking part in a meaningful conversation. McCarthy recognizes that human "hardware" does not support multiple personalities. Physical person and personality are not the same. On the contrary, situations like this are common in computing e.g. time-sharing programs.

Introspection is a key feature that an intelligent machine should have, enabling it to observe its memory and generate introspective analysis of new propositions. An external examiner would assess such ability as consciousness. This contrasts Penrose's belief that AI has no hope of constructing a computer program encapsulating "judgment-forming" ability.

11. Sloman: Architecture vs Algorithm

Sloman argues in favor of AI, recognizing some ambiguity in what was discussed before. He argues that there exist eight different AI theses (Sloman 1992). Penrose attacks only the strongest versions. In fact, following their implied reasoning would imply the existence of a mind in absurd systems. Nonetheless weaker versions of the AI thesis remain both of interest and problematic at the same time.

Sloman's opinion Penrose starts his whole argument from ill-defined conditions. He stresses the importance of *architecture* in artificial intelligence. He writes "what makes something intelligent is not *what* it does, but *how* it does it". This is another significant mistake Penrose makes because he explains mental capabilities and mental states just by developing his argument on the concept of an algorithm.

What's wrong with the strongest AI thesis? The strongest version of the AI thesis fails at describing intelligence because it leads to the conclusion that all sorts of absurd things have a mind, i.e. it is sufficient to possess any computational structure to be deemed as intelligent. This thesis postulates a single stream of processing, but this is not enough to match the complexity that is required for mentality.

Abstractions and static structures can't have minds. Sloman formulates the first AI thesis *T1*. It defines the unknown algorithm that determines intelligence "The Undiscovered Algorithm for Intelligence"(UAI). According to T1, any instantiation of UAI has a mind. Sloman disregards this Strong AI thesis.

Intelligence requires mechanism and activity. T1 might be saved if reformulated in order to include an additional attribute of the instantiation. T1a states that any *temporally* ordered instantiation of UAI will have a mind. But adding the temporal ordering requirement is not enough unless further details and/or constraints are postulated. This new conception of AI is called T2 and it is based on a revision of the term "enaction". It indicates a causal connection between the explicit representation of a program and the structure that is produced. T2 poses now a new challenge, i.e. to specify the sort of causal relation that is sufficient to determine mental states.

Is Turing machine power relevant to intelligence? Human beings and even animals have mental states but there is no evidence that the power of a Turing machine is implied. Therefore Turing machines may not be necessary for mental states.

Moreover, since human beings can make mistakes, this shows that sheer computational power is not a requirement for mentality. This is given by the right functional architecture which relates to how a machine interacts with the external world, or the environment.

Can a single algorithm suffice for intelligence? Discussing serial and parallel computation Sloman stresses the importance of architecture in defining intelligent systems. The key point that should be remarked is that: real parallel computation requires the addition of *transducers* to let the machine interact with the surroundings.

Requirement for a mind. A mind cannot be conceived as a single sequence of states that are ordered in time. The coexistence of perceptual states, beliefs, intentions, moods, etc. that interact concurrently and asynchronously, requires the design of an architecture that is capable of supporting all these intertwined states, and processing them with proper causal links.

Is one processor enough? Sloman argues in favor of the requirement of distributed processors running a collection of computational processes for intelligence. The thesis rules out any engineering objection to the idea of a uniprocessor implementation of the multi-processing architecture needed for intelligence.

Entities with and without minds. It has been established that we might need a very complex and changing architecture running processes simultaneously to explain a mind in terms of a computational process. The processes might serve many different purposes concerned with all human abilities. The right functional architecture may comprise a varying collection of algorithms completing several different tasks.

Is the nature of consciousness self-evident? Sloman's idea is based on the belief that the architecture for human mental capabilities is very complex. Realizing it may imply unraveling many sub-mechanisms that are concerned with different kinds of internal monitoring and control. Sloman argues that "we have only a dim and confused awareness of some of this internal richness that leads people to think they know what they mean by consciousness."(Sloman 1992)

Part 6. Attempts at an answer

12. Searle

A machine can in principle be given the capacity to understand.

The simplest way of putting it is that "we are precisely such machines" (Searle 1980, page 422). If someday we are capable of artificially producing a machine that essentially duplicates the nervous system, then the existence of a thinking machine becomes a simple cause-effect matter. Consciousness and intentionality seem to be prone to implementation by use of different chemistry than the one that is specifically used by us, and it surely is easy to imagine such a scenario.

Instantiating a program is not a sufficient condition for understanding. Formal symbol manipulations are meaningless, not accounting for intentionality. The symbols themselves are not symbols in the true sense of the word, having only "a syntax, but no semantics"(page 422). The Chinese room thought experiment shows that even an intentional system can be equipped with the machinery - the formal program - to produce a series of outputs that carry no intentionality. A belief has no unambiguously formally specified syntactic shape.

A program is independent of where you run it. Mental states, however, are highly dependent on the physical and chemical structure of the specified brain where they are - using the language of programs - instantiated. Thus, one cannot establish an equivalence between the causal powers of the mind and those of a machine running a program. Strong AI, defined as being about programs, not machines, cannot have a hope to describe the very particular type of thinking machine that is the brain.

13. Penrose

Penrose agrees with Searle to the extent that strong AI in itself cannot describe consciousness and intentionality. He speculates about how hypothetical new physics could contribute to understanding mental processes. He does so by arguing that the underlying mechanisms may require a novel sophisticated physical description, in the regime between quantum and classical physics. The problem is that brain processes seem to be non computable in the sense of complexity theory: there are problems that a conventional computer cannot solve, or would take billions of years to do so, while the human brain seems to have the ability to circumvent this limitation.

14. McCarthy and Sloman

McCarthy's idea is that the future advance in computer programs and AI depends on finding a method to make a computer match human thinking in the commonsense informatic situation.

McCarthy stresses the importance of introspective programs, giving computers a sort of memory that they could use to replicate human-like ability to skim through memory when a question is asked.

Sloman's devotes his critic to reasserting the importance of the architecture in defining what is intelligent and suggests the use of parallel computing machines to better simulate human thinking.

Closing Thoughts

The AI approach to understanding the workings of the mind is a recent one. Cognitive science too is a relatively recent, multidisciplinary field. It is not an easy task to build a thinking machine, and it is clear that in spite of the recent successes of AI, which has come a long way since the 1940's, what we currently have are *not* thinking machines. There is in principle no purely logical philosophical impediment to the completion of such a task, and in that sense a machine could think. On the other hand, the nature of consciousness is an unresolved issue that still lacks a scientific explanation.

Whether a thinking machine could come to exist or is only a theoretical abstraction only time will tell. We have clarified some issues on the topic. Crucially, we have distinguished between strong and weak AI, and argued that an explanation for the workings of the mind cannot rely solely on AI. This does not imply that we have no hope of reproducing the inner workings of the mind; it simply means that it will surely be a collaborative effort, and the work being carried out in the fields of cognitive science, philosophy of the mind and of science, together with AI, neuroscience, physics and psychology shall play a role in tackling this huge and elusive challenge.

References

McCarthy, John. 1990. Review of The Emperor's New Mind by Roger Penrose. *Bulletin of the American Mathematical Society* 23 (2): 606–616.

Penrose, Roger. 1989. *The emperor's new mind: concerning computers, minds, and the laws of physics.* Oxford University Press.

———. 1990. Précis of The Emperor's New Mind: concerning computers, minds, and the laws of physics. *Behavioral and Brain Sciences* 13:643–705.

Searle, John. 1980. Minds, Brains, and Programs. *The Behavioral and Brain Sciences* 3:417–457.

———. 2004. *Mind: a brief introduction.* Oxford University Press.

Sloman, Aaron. 1992. The Emperor's Real Mind: review of Roger Penrose's The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics. *Artificial Intelligence* 56:355–396.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind* 49:433–460.