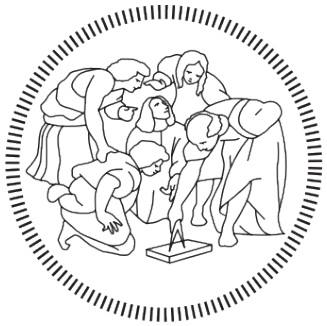
Should machines think?

A perspective on the goals of Artificial Intelligence

Davide Bergamaschi



Politecnico di Milano

2019

1 Introduction

After a long sequence of alternating winters and springs, it is only in recent times that humanity has begun witnessing the fruits of AI research. From self-driving cars to artificial systems substituting humans in more and more complex white-collar jobs, the socio-economic impact of an incessantly increasing level of automation of human tasks is yet to be fully estimated, not to speak about the moral predicaments that come with it.

In these times of exciting innovations and restless technological development, it seems worthwhile to stop for a second and ponder over one fundamental question that is nowadays more relevant than ever: what are the goals of Artificial Intelligence?

The following analysis will revolve around two main approaches in conceptualizing AI - Weak AI and Strong AI. The main ideas behind the two paradigms will be presented, and it will be shown how they naturally tend to pursue different goals. The soundness of the Strong AI paradigm will be questioned, and Weak AI will be defended as a more cautious yet equally effective alternative.

2 Preliminary definitions

2.1 What is AI?

Providing a thorough and universal definition of AI is notoriously a delicate problem which will not be exhaustively addressed here. However the question of *what AI is* is very much tied to the question of what AI *does* and ultimately *should do*, and is therefore deeply connected to the following discussion.

Russell and Norvig (2016) provide a popular concise taxonomy of AI definitions, which are grouped precisely by the kind of goals established by each one. A fundamental distinction is drawn between those who view AI as a quest for reproducing human intelligence artificially and those who identify it with the sheer construction of rational systems.

The authors position themselves in the latter category, and in particular encourage an account of AI focused on rational action (rather than rational thought): AI systems are described as agents attempting to maximize a performance measure on the execution of a certain task, a definition which is conveniently operative yet perfectly general.

It will later be shown in what ways the distinctions outlined above can be related to the concepts of Strong AI and Weak AI, which are about to be defined.

2.2 Weak AI versus Strong AI

In recent years a sort of non-philosophical usage of the terms Weak and Strong AI has emerged, to distinguish among AI that focuses on a specific task and general-purpose AI. Instead, the theoretical dichotomy that that this paper refers to dates back to (Searle, 1980).

Strong AI is introduced by Searle as the position of those who not only believe that computers have the ability to model the human mind, but also maintain that "the appropriately programmed computer really is a mind" (Searle, 1980, 417).

According to Strong AI supporters, it is or it will be possible for AI to build computing systems that have cognitive states entirely similar to the ones produced by the human

brain. Therefore AI can explain human cognition, and it should be regarded as a field of philosophical and psychological research (Dennett, 1978).

Weak AI on the other hand does not advance any claim concerning cognition. Weak AI proponents acknowledge that the "intelligent" systems engineered by AI just act *as if* they were intelligent (Russell and Norvig, 2016), but do not possess any sentience or understanding resembling the human ones. This state of affairs is regarded as unlikely to change in the future, at least for what concerns the traditional tools of AI (computers and programs).

This belief however is not perceived as limiting of the innovative potential of AI research, nor it is meant to deny that computers can be a powerful tool in the study of the mind (Searle, 1980).

2.3 Behaviorists and computationalists

To arrange more clearly the critique of Strong AI that will be delivered later, it is useful to characterize two main subtrends of it which ground their claims on deeply different foundations: Behaviorist Strong AI (BSAI) and Computationalist Strong AI (CSAI).

The term BSAI is meant to encapsulate the tendency of attributing mental activity to AI artifacts on the base of their display of rational behavior.

The core claim of BSAI supporters is that if an artificial system exhibits a certain level of general rationality or performs one or more highly complex tasks that are generally regarded as requiring human understanding, then that artificial system must possess real mental capabilities.

The term CSAI refers instead to the Strong AI positions that stem from computational theories of the mind. The common denominator of this trend is the belief that the brain is a computing system, and that all is needed for cognition is implementing its relevant computation.

For CSAI supporters, the instantiation of a specific program on a computer possessing a sufficient amount of computational resources is enough to give rise to a mind capable of understanding and possibly experiencing reality.

3 A matter of goals

It has to be noted how a shift between the Weak AI and Strong AI position can really influence the overall mindset of an AI practitioner.

If the claims of Strong AI are embraced, the most natural goal of AI would arguably become the one "to create artificial persons: machines that have all the mental powers we have, including phenomenal consciousness" (Bringsjord and Govindarajulu, 2018, 8.1). Note how this type of inclination makes Strong AI particularly compatible with a definition of AI focused on an anthropomorphic conception of intelligence.

As it will be argued, it is far from obvious that AI as a branch of computer science will have a major role in carrying out the ambitious undertaking outlined above. And even if it was theoretically possible to make computers think and feel like humans, the concrete possibility of doing so will in all likelihood come very far in the future.

On the contrary, Weak AI settles for the goal of simulating intelligence, finding more and more effective ways to capture intelligent behavior in mechanical procedures, regardless of how the human brain works.

In light of the above, some argue that the ultimate aim of the Weak AI paradigm can be identified with constructing an artificial system capable of passing for a human in all behaviors (Bringsjord and Govindarajulu, 2018).

However it should be noted that the aim of Weak AI can clearly be super-human (rather than human-like) performance, since AI already surpasses human experts in a number of applications (e.g. playing chess). Increasing that number, with particular attention to those fields that could directly improve human life, seems an equally valid, and maybe more realistic, goal for Weak AI.

It is clear how this view can be naturally combined with the definition of AI suggested by Russel and Norvig: Weak AI can be ultimately seen as the creation of rational agents which automatically and effectively accomplish a task better than humans, with a focus on performance rather than anthropomorphism.

Notice how anyway the specificity of the domain of application of the agents built up to a certain point will not prevent the future creation of an entire human-like system, nor does it obstruct the quest for a major achievement such as Artificial General Intelligence: both of the two undertakings can surely benefit from the know-how gained on individual specific applications.

4 Debating Strong AI

While the Weak AI position, probably by virtue of the modest nature of its claims, has not subjected to specific philosophical attacks [1], several criticisms of Strong AI have been raised throughout the years.

Drawing from past debate, the following section aims at presenting some arguments and intuitions which can counter the Strong AI perspective.

4.1 Addressing BSAI

A strong behaviorist trend can be traced back to the very roots of AI. In (Turing, 1950), a classic paper as well as the philosophical cornerstone of the BSAI perspective, Alan Turing attempts to address a fundamental question: "Can machines think?".

Noting from the very beginning that the question is ill-posed, he proposes to replace it with an operative test, now widely known as the Turing Test (TT), to verify whether a machine M is able to think: given a human interrogator N and a human participant H, can N, after asking H and M a series of questions in natural language, tell who among them is a machine? If N is not able do so with a high success rate, due to M being able to answer general questions in a very convincing way, then one must concede, in Turing's opinion, that M is indeed able to think.

A straightforward way to counter the soundness of the TT is to provide a suitable counterexample, that is to show a theoretically feasible system that can pass the TT but evidently is not capable of general thought.

To achieve this, let's imagine that there exists a program P that can pass the TT. Without dwelling too much on the definition of program, P can be seen with enough generality

[1] Some general arguments against AI also apply to Weak AI, but do not seem to have particular force. An example is (Penrose et al., 1989), proposing a famous as well as controversial Gödelian argument (effectively refuted by Searle, 1998, among others).

as a procedure which takes questions as input and produces proper replies as output in a deterministic fashion. Assuming, just like in the original TT, that the conversation consists in a finite amount of finite exchanges, it is theoretically possible to enumerate all of the input-output bindings produced by P. One can then conceive without too much trouble a program P' that perfectly reproduces the behavior of P using some trivial mechanism.

For example, similarly to what has been noted by Block (1981) among others, it would be theoretically possible to build a huge lookup-table matching each possible question q and each possible conversation history up to q to the proper answer a, with q and a being shorter than an arbitrarily large bound L, sufficient for the purposes of the test. A program P' could then just keep track of the conversation history and perform matches in the lookup-table to give intelligent answers to the interrogator.

An instantiation of P' would then, by definition, pass the TT, but even the most committed behaviorist would probably hesitate before saying that such an instantiation is thinking in a meaningful way during the test. Furthermore, it is clear that despite responding properly to general dialogue, P' does not capture general intelligence, since for example it is not equipped to answer questions longer than L, nor it can learn to do so.

Critics who point out that the lookup-table needed for P' would be so huge that it would never be physically realizable in this universe are most probably right, but they overlook the fact that practical feasibility is not needed for this type of counter-argument. The sheer theoretical possibility of the existence of a trivial program that could pass the TT undermines its sufficiency to detect "real" intelligence.

Notice how strictly behavioral extensions of the Turing Test, e.g. requiring sensorimotor capabilities as in the Total Turing Test (Harnad, 1991), are still vulnerable to the type of critique that has been presented above. For example, the necessity of adding robotic parts to machine M to make it physically act like a human does not pose any condition on the triviality of the program controlling M's parts.

Another philosophical blow to the BSAI perspective has been struck by John Searle. In (Searle, 1980) he proposed a famous *Gedankenexperiment*, now known as The Chinese Room, attempting to demonstrate how even a possibly elaborate program which accomplishes a complex task cannot produce mental activity.

To briefly summarize it, the experiment goes as follows. A man is locked in a room. From the outside, questions in Chinese are sent in. The man does not understand Chinese at all, but has at his disposal a rule book containing exhaustive instructions on how to correlate input Chinese characters with output Chinese characters, without any further semantic explanation. All the man does then is blindly following such instructions and sending out the obtained sequences of symbols, meaningless squiggles to him but perfectly intelligible answers for the Chinese speakers outside the room.

The analogy is clear: the man in the room behaves exactly like a computer executing a program (the one specified by the rule ledger). From the outside, the system certainly appears to understand Chinese. However, once its inner workings are disclosed, one can intuitively conclude, Searle argues, that it does not. The intuition is driven by the fact that the man, the only intelligent element in the room, does not acquire any understanding of the Chinese language whatsoever by going through all the calculations.

In particular, no matter what program is fed to a computer, it will never have the same understanding of the human mind: minds have semantic contents, but the kind of manipulation performed by a computer is purely syntactic and syntax by itself is not sufficient for semantics (Searle, 1998). The possibility of semantics typical of the human mind is

guaranteed, according to Searle, by a set of yet unknown biological features of the brain, which a computer by itself lacks.

One may or may not find Searle's argument compelling. Many for example respond that at the low-level the brain operates with electric signals devoid of semantics as well, and still the human mind emerges from it.

Nonetheless it should be recognized how the Chinese Room experiment is effective at suggesting a crucial point: the possibility of intelligent behavior by an appropriately programmed computer and the idea that a computer supposedly shares with the human brain the characteristic of being a syntactic system are *not* sufficient reasons to conclude that computers and brains can have the same type of cognition.

Strong AI supporters are left with the burdensome task of specifying how and why a human-like mind would arise from symbol manipulation, regardless of the physical substratum where the computation is instantiated. BSAI is clearly insufficient to provide an answer to those questions, while, as it will be shown, CSAI precisely aims at doing so.

4.2 Addressing CSAI

There are at least two interesting objections to Searle's experiment. The "systems reply" maintains that, while the single components of the Chinese room do not have understanding, the system as a whole may do. The "brain simulator reply" instead proposes to replace the generic program with one simulating the workings of the brain of an actual Chinese speaker, so as to generate real understanding.

A sort of combination of the two replies brings up some non-trivial questions. What if the human mind could be described entirely in computational terms? What if the instantiation of some relevant computation is enough to *implement* an actual mind, rather than just simulate it?

It is not hard to see why computational accounts of the human mind have been very popular in cognitive science throughout the years. The brain naturally appears to be the information-processing organ *par excellence*, receiving sensory input from the environment, reasoning, and producing suitable motory output as a response, which makes "the computer metaphor" undeniably tempting.

Several computationalist theories have therefore been proposed, different in their peculiar definition of what computation is and how it can be realized by a physical system, but united by the fundamental idea that the mind *is* a computing system, rather than just resembling one.

Analyzing said theories in depth goes well beyond the scope of this work. Instead, reference will be made mainly to the version of computationalism put forward by (Chalmers, 2011), since it does not advance too bold empirical hypotheses and is still representative of the core CSAI claims. The aim here is not to provide a thorough refutation, but rather highlight some critical aspects and suggest the possibility an alternative perspective.

Chalmers defines computation as an abstract specification of the causal structure of a system, describing the pattern of interaction among the parts of the system without fully going into physical details. A physical system can be then said to implement a computation if its causal structure mirrors the formal structure of the computation.

The generality of the notion of computation employed by CSAI proponents has often been under the attack of anti-computationalists, who observe that most natural processes can

be described in computational terms, making computationalism trivial (Rescorla, 2017). Computationalists rebut that cognition has a special bond with computation, supporting the core assumption that "it is in virtue of implementing some computation that a system is cognitive" (Chalmers, 2011, 2.2). That is, differently from a biological process like digestion, implementing the right computation actually produces cognition, rather than just simulating it, even if no brain is involved.

Various attempts have been made to justify such a strong claim. Chalmers does it by heavily relying on a functionalist view of the mind (Rescorla, 2017). He claims that mental properties are organizational invariants in the causal topology of a cognitive system, that is two systems with the same causal organization must share the same mental states. A computational model would then be a most natural way to spell out the relevant causal topology of a cognitive system, which could be theoretically realized on a multitude of physical systems (including a digital computer with sufficient resources).

A particularly delicate issue is raised, unsurprisingly, by the presence of consciousness. Chalmers divides mental properties in psychological properties, which are concerned with what the mind does, and phenomenal properties, which are concerned with how the mind feels. While it is assumed that the former can be defined in relation to their causal roles, and are therefore organizational invariants by definition, Chalmers concedes that the functionalist paradigm has troubles explaining the latter.

And indeed, as pointed out by Block (1980) among others, one can imagine a functional equivalent of a person, that has precisely the same psychological states but experiences reality in a significantly different way (for example, by having an inverted "color spectrum").

To address the explanatory inadequacy of functionalism, Chalmers (1995) provides an *ad hoc* thought experiment to show that phenomenal properties too are organizational invariants. In short, the reader is asked to imagine a hypothetical surgical operation where the brain of the patient, who remains conscious throughout the experiment, is gradually substituted with functionally equivalent artificial components (say, for example, neuron by neuron replacement with silicon chips).

Chalmers argues, by appealing to scientific intuition, that it is completely unlikely that consciousness will suddenly disappear in a discontinuous fashion.

Instead one may be tempted to say that consciousness would gradually fade away. In that case however there must be some intermediate state in which the system is partially artificial and partially biological. In such a state, the patient's conscious experiences would be degraded, but, since the substituted component are perfect functional equivalents by hypothesis, no dissonance in the behavior and beliefs of the patient can be present. Since, Chalmers argues, it is implausible that a fully rational being would be significantly wrong about its experiences, it must be that the patient's consciousness is unaffected.

A third possibility would be that the artificial brain had different phenomenal properties than the human one. In this case Chalmers asks to consider two intermediate systems created in the process, whose conscious experiences differ significantly. For example, one may see red where the other sees blue. One could imagine to install alongside the relevant neural circuit a causally isomorphic artificial replacement, with a switch connecting the two. By flipping the switch one could make "red and blue experiences 'dance' before the system's inner eye" (Chalmers, 2011, 3.2), while the patient, as in the above case, could not react or notice. Chalmers concludes that this is equally implausible.

Leaving out comments on the physical feasibility of the experiment, it is immediately clear how the whole argument heavily relies on intuition, despite Chalmers himself acknowledges

that "intuition is unreliable as a guide to empirical possibility" (Chalmers, 1995, 2.) especially when it comes to consciousness.

At the very start, the possibility of a sudden disappearance of consciousness is hurriedly ruled out by appealing to a supposed impossibility to "compound continuity into discontinuity" in the laws of physics. However, as noted by Aides (2015), as a trivial counterexample one could think of the molecule-by-molecule removal of matter from the center of a tense rope, causing its abrupt breakage at some arbitrary point in time.

A possibly more serious fallacy resides in the successive part of the argument. As highlighted in (Van Heuveln et al., 1998), Chalmers takes for granted the identity of the patient and overlooks the possibility that by flipping between two states, two different individuals may be "drifting in and out of their own phenomenal world" (Van Heuveln et al., 1998, 4.). There are no guarantees that new individual would be able to faithfully recollect experiences from the old individual, so as to notice that something has changed (for example, his ability to internally visualize the color red may be functionally impaired). Even if the changes in perception after a replacement were huge, there might just be no inner eye to tell the difference.

Notice how these critical points do not even question the overall validity of a functional description of psychological states, which raises concerns even among CSAI supporters and would deserve a thorough critique in its own right (see Rescorla, 2017, for a list of critiques).

All of the above shows how the theoretical foundations of even one of the most well-grounded (and minimal) computational theory are not as solid as they are claimed to be, and appeal to intuition no less than the positions of their philosophical adversaries.

But the weakness of the computationalist description lies not only in its premises but also in its conclusions: what is presented as the most plausible theoretical possibility, brings rather implausible consequences. Block (1980) imagines a scenario where the whole population of China is suitably organized to implement the relevant computation performed by neurons. Can one lightheartedly accept that this will give rise to a group mind, complete with phenomenal consciousness, as most CSAI theories would suggest? Some CSAI scholars such as Chalmers (1996) go as far as to argue that, since the bond between cognition and computation is so tight, even an extremely simple information-processing system like thermostat may have some kind of phenomenal consciousness, paving the way for a sort of panpsychism.

It is clear that neither side has the final word in the CSAI discussion, which at some level strays into the realm of belief rather than staying in the one of rigorous demonstration. While their critics continue being skeptical, CSAI supporters keep believing that computers will be able to think just like humans. However, the burden of proof of such an extraordinary claim still falls onto the latter. Neuroscience appears to be the only voice authoritative enough to break this philosophical stand-off, providing definitive answers on the inner workings of the brain. Yet, it is unclear whether this will happen in the near future or in several centuries from now. In the meantime, a cautious approach such as the one of Weak AI seems the most reasonable one.

5 Conclusions

Some major philosophical flaws of Behavioral Strong AI have been exhibited, and it has been shown how Computationalist Strong AI fails to provide any conclusive proof for its claims. Weak AI has been proposed as an alternative paradigm which is more cautious

and at the same time fully capable of bringing around the life-changing innovations that humanity expects from AI.

It is entirely possible that future scientific research will discover that cognition and computation are not as inextricably bound as some had thought, and that computers and programs are not sufficient by themselves to fulfill the Promethean dream of constructing an artificial person.

However, the above scenario would not constitute a defeat for AI at all. AI research will continue raising the bar of machine intelligence regardless, effectively automating an increasingly large pool of human activities, and playing a central role in future technological innovation. The fact that machines will or will not be capable of really thinking will remain a matter of secondary importance in this advancement.

References

Aides, N. (2015), 'Yet another objection to fading and dancing qualia'.

Block, N. (1980), 'Troubles with functionalism', *Readings in philosophy of psychology* **1**, 268–305.

Block, N. (1981), 'Psychologism and behaviorism', *The Philosophical Review* **90**(1), 5–43.

Bringsjord, S. and Govindarajulu, N. S. (2018), Artificial intelligence, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', fall 2018 edn, Metaphysics Research Lab, Stanford University.

Chalmers, D. J. (1995), 'Absent qualia, fading qualia, dancing qualia', *Conscious experience* pp. 309–328.

Chalmers, D. J. (1996), *The conscious mind: In search of a fundamental theory*, Oxford university press.

Chalmers, D. J. (2011), 'A computational foundation for the study of cognition', *Journal of Cognitive Science* **12**(4), 323–357.

Dennett, D. C. (1978), 'Artificial intelligence as philosophy and as psychology', *Brainstorms: Philosophical essays on mind and psychology* pp. 109–26.

Harnad, S. (1991), 'Other bodies, other minds: A machine incarnation of an old philosophical problem', *Minds and Machines* **1**(1), 43–54.

Penrose, R., Penrose, R., Gardner, M. and Press, O. U. (1989), *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, Oxford landmark science, Oxford University Press.

Rescorla, M. (2017), The computational theory of mind, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', spring 2017 edn, Metaphysics Research Lab, Stanford University.

Russell, S. and Norvig, P. (2016), *Artificial Intelligence: A Modern Approach*, Always learning, Pearson.

Searle, J. R. (1980), 'Minds, brains, and programs', *Behavioral and brain sciences* **3**(3), 417–424.

Searle, J. R. (1998), *The Mystery of Consciousness*, Granta Books.

Turing, A. M. (1950), 'Computing machinery and intelligence', *Mind* **59**(236), 433–460.

Van Heuveln, B., Dietrich, E. and Oshima, M. (1998), 'Let's dance! the equivocation in chalmers' dancing qualia argument', *Minds and Machines* **8**(2), 237–249.