Abstract

The Computation Theory of Mind (CTM) is the thesis that mind is a computational system. It proved to be a productive hypothesis in the area of artificial intelligence, since various cognitive tasks have found a computational explanation and implementation. Some questions resisted though. In particular, CTM seems to provide no insight into the nature of such an important mental phenomenon as consciousness. This was used as a major criticism of CTM as a general theory of mind. In this paper I explain Steven Pinker’s take on the problem of consciousness and put it in the context of the modern CTM approaches. In a nutshell, Steven Pinker accepts the usual division between easy and hard (unsolvable) problems of consciousness, whereas the contemporary research tries to go beyond this distinction.

1 The Computational Theory of Mind

The Computation Theory of Mind (CTM) is the thesis that mind is a computational system. Since there are many ways in which “mind as computational system” can be understood, CTM corresponds to a large family of views rather than a unique theory. For a review of different flavours of CTM I suggest having a look at [16]. From there I borrow the most general definition of CTM, which nevertheless captures the distinct nature of CTM relative to other theories of mind: CTM is a thesis that cognition is Turing-style computation over mental symbols.

Computer simulations are widely used in other cognitive science research programmes that see themselves as rivals of CTM: namely, in computational
neuroscience, type-physicalism and embodied cognition\(^1\): so what does distinguish CTM from other theories?

To answer this question it suffices to note that if cognition is understood as Turing-style computation over mental symbols, then CTM implies Multiple Realizability and Internal Representations theses. The aforementioned theories deny at least one of these consequences.

**Multiple Realizability** It is uncontroversial that computation as understood in computer science can be implemented via different media. Thus, if mind is Turing-style computation, it can be realized not only in brain, but also using, for instance, silicon chips. Clearly, for Artificial Intelligence only cognitive science theories that imply Multiple Realizability are of interest. Later we will see that in relation to consciousness, the rejection of Multiple Realizability thesis is one of the main argument strategies against CTM.

**Internal Representations** Since CTM equates mind with symbolic computation, but rejects dualism (i.e., it doesn’t distinguish brain and mind), it must hold that in some way brain contains “mental symbols”, that is, it has structures that correspond to the symbols on which computation is done. Usually these symbols are called internal representations.

Computational neuroscience and type-physicalism deny Multiple Realizability thesis, although are often sympathetic towards Internal Representations (cf. [15]), whereas embodied cognition theorists accept Multiple Realizability, but strongly argue against existence or importance of internal representations.

CTM is supposed to be a universal theory of mind: it should be possible to explain every aspect of the mind by providing a suitable computational description and then, by Multiple Realizability, it should be theoretically possible to reproduce it artificially. This explains why question of consciousness is important for CTM: intuitively, it points at its limits. If CTM is true, then by Multiple realizability it should be possible to create artificial machines that are conscious. But this is a consequence that many find hard

---

\(^1\)Computational neuroscience is a branch of neuroscience that uses computer simulations to study brain; it is not a theory of mind, but a method of studying the brain. Type-physicalism is a view that the cognitive properties can be only explained by the specific physical and chemical makings of the brain. Embodied cognition is a theory that emphasizes the role of the environment in cognition and denies that the mind exists separately from the environment. Stanford Encyclopedia of Philosophy is a great place to get an overview of these theories: plato.stanford.edu.
to accept. Thus to defend their theory CTM proponents have to provide an explanation of how such machines are possible. In this paper I present Steven Pinker’s view on computational theory of consciousness, which can be seen as paradigmatic of classical CTM (cf. [16]), and discuss how the developments on the topic made since he wrote his book can be placed in the theoretical framework he used.

2 Steven Pinker on Consciousness

Pinker discusses the computational approach to consciousness in Chapter 2 (pp. 131–148) and Chapter 8 (pp. 561–565) of the book “How the Mind Works?” [13]. In the paper he wrote a decade later he makes the same arguments [14].

First of all, it is important to understand what exactly we want to explain. To this end Pinker makes distinction between three possible understandings of the concept of consciousness; he will only be interested in two of them. I suggest looking at his definitions in terms of the famous distinction between easy and hard problems of consciousness made earlier by Chalmers [2].

2.1 Easy and Hard Problem of Consciousness

Pinker and most of the theoretical investigations on consciousness after him in some way or another use the distinction between easy and hard problems made by Chalmers. Intuitively, the easy problems of consciousness are those that seem to be directly susceptible to the explanations in terms of computational or neural mechanisms. The hard problems are those that seem to resist the methods of cognitive science.

The easy problems of consciousness include those of explaining the following phenomena:

- the ability to discriminate, categorize, and react to environmental stimuli;
- the integration of information by a cognitive system;
- the reportability of mental states;
- the ability of a system to access its own internal states;
- the focus of attention;
• the deliberate control of behavior;
• the difference between wakefulness and sleep.

The hard problem consists of explaining how subjective experience arises from neural computation:

It is undeniable that some organisms are subjects of experience. But the question of how it is that these systems are subjects of experience is perplexing. Why is it that when our cognitive systems engage in visual and auditory information-processing, we have visual or auditory experience: the quality of deep blue, the sensation of middle C? How can we explain why there is something it is like to entertain a mental image, or to experience an emotion? It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises. Why should physical processing give rise to a rich inner life at all? It seems objectively unreasonable that it should, and yet it does [2].

For instance, in the realm of color perception we can hope to solve the following easy questions: how do we distinguish green from red? how do we form a concept “green” in the natural languages? how does it relate to the way we perceive green? The hard problem would be to explain the very experience of “greenness”.

Note that according to Chalmers the problems are easy or hard not only for CTM, but any cognitive science theory. Critics of CTM would say that for their favorite theory this distinction doesn’t apply, that is, for a successful theory of mind all problems are and should be easy [17]. It may also be useful to note that for CTM, by Multiple Realizability, we automatically get easy AI problems and hard AI problems: create a machine that runs computations that correspond to easy and hard problems respectively.

2.2 Three Definitions of Consciousness

Pinker distinguishes three meanings of the concept of consciousness: self-knowledge, access to information and sentience. The first two turn out to be easy, whereas the last one corresponds to the hard problem of consciousness as defined above.
Self-knowledge  Typically when we say consciousness we mean the ability of an intelligent being to have information about itself, in other words, self-awareness. This understanding of the concept justifies one of the most advertised “tests for consciousness” in cognitive science: if an animal recognizes itself in the mirror it may be considered conscious [7]. Pinker is not interested in consciousness as self-knowledge, since he argues that it doesn’t constitute a separate hard problem. Self-knowledge is implied by explanation of knowledge in general: one can easily make a robot that recognizes and analyses itself, once it can recognize anything else.

Access to information  Sometimes we speak of consciousness when we want to distinguish between information that is readily available to us and information about inner workings of our mind and body that is hidden. For instance, we are not conscious of the computations in the brain that recover 3-D shapes from 2-D retinas, but we are conscious of the content of what we are looking at, such as shape or color. Although this problem seems to be more interesting than consciousness as self-knowledge, Pinker will argue that it is still an easy question and not what we are really interested in when we talk about the mystery of consciousness [14]. A simple computer metaphor provides an intuition about how the explanation and implementation of consciousness as access to information could go. If you cannot print a page from your computer that is connected to a printer, you don’t necessarily know why this is the case and might have to check the printer to find what the problem is. Your computer has access to the information about whether the printer is working or not, but not necessarily to the reasons why it is not working properly; in contrast, the printer itself may contain a message telling you what is wrong (“Place paper in tray 2”), since this information is available to the printing machine.

Sentience  The meaning of consciousness Pinker is most interested in corresponds to the subjective or phenomenal experience, in philosophical tradition also known as qualia [19]. Here the theory of mind should explain how the experience of “redness” comes into being (and for CTM it means to show how the experience of “redness” is computed), in contrast to plain understanding of the concept or its correct usage. This is the sense in which consciousness is the hard problem for CTM.
2.3 Consciousness as Access to Information

According to Pinker, we are quite close to solving the easy problem for consciousness – consciousness as access to information – since its main features: *sensory awareness, focal attention, emotional coloring* and *the will* – all can be explained within CTM. Namely, the design specification of any efficient information processor operating in the real world can explain why access-consciousness exists and how it can be possibly implemented.

The main observation is that humans are not able store all the information needed to perform tasks of any more or less complex task, due to our physical limitations and the nature of information. For instance, the amount of possible sentences in English or a chess game significantly exceeds the size, time and resources available to the brain. Therefore, some of the information has to be derived. What should be stored are the rules by which a correct derivation can be made on demand, that is, consciously. In order to make such derivations efficient only relevant information should be used and thus some information is hidden from the subject. Then the main features of consciousness-as-access are then explained in terms of efficiency of information processing. This view is currently developed in detail under the title *Global Workspace Theory* [6].

**Sensory awareness** It can be argued that the most computationally beneficial for perception is “intermediate-level consciousness” that we possess [9]: to perform the most wide variety of tasks it is better not to concentrate on particular sounds or shades.

**Focal attention** Similarly, simple recognition tasks can be performed without conscious deliberation or attention: as experiments in cognitive science show, only when we need to perform non-mundane tasks that our attention is needed (cf. [18]).

**Emotional coloring** Some thoughts and perceptions are accompanied by emotions. This is needed to trigger goal states for the relevant computation; the presence of particular emotions in certain situations can then be explained evolutionary.

**Will** From computational point of view “I” that makes decisions can be understood as “another set of if-then rules” that we need in order to perform the multitude of tasks coherently and in an efficient manner.
At this point one might ask: if we explained all these features, haven’t we already explained consciousness? Pinker holds that consciousness as access to information doesn’t provide a solution to the problem of consciousness as sentience and thus this explanation is not complete. The reason for that is that the solution to the hard AI problem doesn’t seem to follow from these easy explanations, as shown by the Chinese room mental experiment.

2.4 Consciousness as Sentience and the Chinese Room

If we claim to solve the hard problem of consciousness within CTM, then in terms of the hard AI problem we would get a positive answer to the following question: “If we ever duplicate the information processing in the human mind as an enormous computer program, would a computer running the program be conscious?” By the Chinese room argument, which is one of the most famous arguments against CTM made by John Searle, this cannot happen.

**Chinese Room Mental Experiment [17]** A person who knows no Chinese is put in a room. Every now and then the pieces of paper with Chinese symbols are slipped under the door. The person in the room follows the list of instructions: when you see [squiggle] [squiggle] [squiggle], write down [squoggle] [squoggle] [squoggle] and slip the paper under the door. The set of instructions is an artificial intelligence program for conversing in Chinese, and the people outside are convinced that they are speaking with the native Chinese speaker. The question is: does the person in the room actually understand Chinese?²

²The picture is taken from this site: http://debategraph.org.
If understanding is running a suitable computer program, as CTM claims, the person who is the room who runs the program does understand Chinese. But by assumption, it is false. Thus, understanding must be different from symbol manipulation or computation. One can say that it is not the person, but the room that performs the computation. But similarly, we assume that the room doesn’t really understand what it is doing, so the argument is supposed to go through in both cases. The argument about impossibility of computational explanation of consciousness is just a variant of the Chinese room experiment, where instead of a Chinese language program the Chinese room runs a consciousness program.

One way of answering to this challenge is to bite the bullet and accept that the Chinese room does understand Chinese or is in fact conscious, contrary to what we intuitively think. Our intuition may be just wrong about the matter. According to Daniel Dennett, for instance, consciousness is a cognitive illusion that doesn’t require an explanation that goes beyond CTM [5].

Patricia and Peter Churchlands provide a mental experiment Luminous Chamber to argue ad absurdum that the reasoning used in the Chinese room
argument could be used to disprove Maxwell’s theory of light as electromagnetic waves [3] [13] (p.93–94).

Imagine a person who waves a magnet. Some electromagnetic waves are created, however, there is no light coming out. If we were to use the same line of thought as in the Chinese room argument, we would have to conclude that light is not caused by electromagnetic waves. But we know that rapid waves would cause light to appear. Similarly, if we have fast and intricate enough computations, it is not so clear whether we can claim that they don’t produce understanding or consciousness, since our intuitions are based on the simple cases where a person or the Chinese room performs the computation.

Of course, this counter-argument doesn’t prevent one from appealing to philosophical zombies [10]: it is a logical possibility that even if artificial agents perform all the necessary “rapid” computations needed and produce the mental abilities and behaviors people have, they nevertheless possess no consciousness. But by this argument we can deny consciousness to fellow human beings, since there is no proof that the person sitting near you is not a philosophical zombie. This contradicts the basic assumption we make when we talk about consciousness – under normal circumstances humans have it, and that’s what we want to understand or reproduce. In its more sophisticated versions the Chinese Room argument can be seen as an instance of the philosophical zombie argument and thus doesn’t pose a real threat to CTM.

Steven Pinker learns a different lesson from the Chinese Room experiment: since we cannot even tell whether our fellow human beings are conscious, the question of whether actual philosophical zombies can exist is unverifiable. Inasmuch as he doesn’t deny the existence of consciousness as Dennett does, in his version of CTM we are left with no explanation of what consciousness is and how it is related to solvable problems. In the end he speculates that explaining sentience just goes beyond our cognitive abilities as a species – we cannot relate mechanistic explanations with our experience in the same that monkey cannot learn long division.

3 Again, I suggest consulting SEP for an overview of different version of the argument [4].
3 Recent Computational Theories of Consciousness

Although the Chinese room and philosophical zombie arguments may appear abstract, it seems that they correspond well to intuitions that people who are working in artificial intelligence and cognitive science have: as an informal survey of American AI Society in [12] shows, for the majority the hard problem of consciousness is either a pseudo-problem or they think the field is not advanced enough to solve it. Here I briefly summarize the approaches taken in the contemporary literature on computational consciousness and mention how they relate to Pinker’s views.

Self-Model Theory

Although for Pinker self-knowledge was irrelevant with respect to the hard question of consciousness, it is not so for many researchers nowadays. Based on reviews of the recent research in machine consciousness in [12] and [8], the common position of contemporary CTM theorists can be summarized in the following way: Phenomenal consciousness is the property of a computational system X has if X models itself as experiencing things. Under this view consciousness becomes a part of the mechanism of introspection about perception. In [1] it is called “the debugging mode”: mind uses conscious introspection of the internal model to learn faster and to modify its behavior. [12] claims that such an understanding of consciousness also permits to explain qualia: qualia have a function of stopping the introspection – when we reach them, the debugging cannot go further. Here we see interesting attempts to link easy problems to the hard problem of consciousness within one theory, something Pinker thought was conceptually impossible.

Global Workspace Theory

The Global Workspace Theory [6], which seems to be quite established now, may be seen as following Dennett’s view: the hard problem reduces to easy problems. Here one is interested in finding neural correlates for conscious and unconscious information processing and then create computational models based on these findings. Here one doesn’t try to explain qualia in any way. The view corresponds to what Pinker describes as consciousness as access to information.

Information Integration Theory

Information Integration Theory claims that there is a quantifiable measure of consciousness. Since some very sim-

\[\text{See } \text{http://plato.stanford.edu/entries/consciousness/#InfIntThe}\]
ple systems have the identified measure, to test whether the system posses consciousness we need an additional behavioural test. Here the test should show whether in the agent’s artificial mind the information is integrated in a human-like way. For a rough example, take two pictures where what is depicted on one of them somehow breaks a cultural pattern (instead of a keyboard there is plant in front of a computer, for instance) and ask the question: which picture is wrong? From the point of view of Information Integration Theory, the machine that can answer such a question is characterized by highly integrated information processing characteristic of human mind [11]. Of course, this view is agnostic about whether we could solve the hard AI problem and actually build such an agent. For instance, based on analysis of recent neuroscience developments [15] argues that temporal and information-density dynamics that are defining features of human mental states cannot be realized outside brain.

Thus we see that current researchers in computational consciousness avoid Pinker’s pessimistic conclusion in various ways. Information Integration Theory plainly ignores the unverifiability argument and traditionally relies on the Turing test. Global Workspace Theory rejects the distinction between easy and hard problems and concentrates on consciousness as access to information. Finally, Self-Model Theory aims to falsify Pinker’s claim that it is impossible to bridge easy and hard problems of consciousness and therefore leaves some hope that CTM can eventually explain the mystery of consciousness.

References


