On computational conceptions of mind and brain

Max Pohlmann

1st June 2022

(This essay was written as part of the assessment of the course *Rationality, Cognition and Reasoning*, taught by Dr. Levin Hornischer at the University of Amsterdam.)

1 Introduction

The conception of our minds or brains in computational terms is pervasive nowadays. This can take many forms, and the distinction between mind and brain is not always clear. Claims one often reads in the cognitive science literature are not unlike the following: 'the brain is a digital computer'; 'the mind is software running on the hardware that is the brain'; 'the brain is (/can be modelled as) a Turing machine'. The conceptions underlying these claims fall under the umbrella of what is known as *computational theory of mind* or *computationalism*. Indeed, a computationalist conception of mind/brain constitutes the basis of most endeavours in cognitive science: in order to understand an aspect of cognition, it is assumed that the brain implements a function/algorithm that the researcher then aims to replicate. Van Rooij [vR08], for example, takes this conception more or less as given and claims that the brain can only implement functions that are tractable in some sense.

What I want to do in this essay is consider challenges to this computationalist view and see to what extent it is tenable as a basis for a science of the mind. Firstly, we shall disentangle some beliefs that might be lumped under the rubric of computationalism, but do not necessarily entail each other. Then, we shall briefly consider some criticisms of computationalism that I belief to be misconceived, whereafter we can move on to more serious objections, which fall into three categories: doubts about computationalist conceptions of the mind altogether, problems with the relationship between mind and brain, and claims that the brain is computationally more powerful than classic computers. I conclude with what I believe to be a tenable limited form of computationalism.

2 Disentangling beliefs

It is readily seen that the claims I listed in the introduction are related, but substantially different. Before we can consider criticisms, we have to demarcate the different beliefs to be criticised.

Searle distinguishes between three questions, which I quote from [Sea90, p. 21]:

1. *Is the brain a digital computer?*
2. *Is the mind a computer program?*
3. *Can the operations of the brain be simulated on a digital computer?*

He takes the first question to be equivalent to ‘Are brain processes computational?’ and subsumes under it the question of whether the brain does information processing. For the purposes of this essay, I shall reformulate his questions as claims and append some essentially equivalent ones to them:

1. The brain is a digital computer. Brain processes are computational. The brain does information processing.
2. The mind is a computer program. The mind is software running on the hardware that is the brain.
3. The (operations of the) brain can be simulated on a digital computer. The brain can be modelled as a Turing machine. The brain is at most as computationally powerful as a Turing machine.

3 Misconceived objections

Before we can move to serious criticisms of the above beliefs, we shall do away with some misconceived objections.

Various results in behavioural psychology, like the Wason selection task [Was68] and the results of Kahneman and Tversky [Kah11], have shown that humans do not follow normative rules and make mistakes, unlike (idealised) computers. It is true that humans do not always seem to follow what might be called normative rules. This does not mean, however, that humans make random mistakes that would not be made by a computer, but can be explained by stating that the algorithm the human (supposedly) executes does not adhere to the given norms. For example, Stenning and van Lambalgen [SvL12] have proposed formalisms that can account for allegedly faulty reasoning done by humans.

The brain has no designated central processing unit or memory banks. It is entirely unlike any modern computer. The widely-accepted Church–Turing thesis tells us that the concept of computation is independent of any particular formalism. Both the lambda calculus and cellular automata are entirely unlike modern computers, but are equivalent in their computational power to Turing machines (which form the theoretical basis of modern computers). The claim is not that the brain works just like a modern computer, but that it is an equivalent implementation, i.e. it somehow performs computation.

With each novel technology, like hydraulic pumps or gearboxes, people tend to think of the brain in such terms. Brain-as-computer is just the newest trendy metaphor.[1] What is different about the computation metaphor is that it abstractly encompasses the previous metaphors. Just as in the previous paragraph, the point was never to claim that there are literal gears in the brain, but that something analogous was going on. This something is, I would say, computation.

Turing machines map inputs to outputs, whereas humans interact continually with the outside world. This is not so much a misconception as pointing out an inaccuracy in the computationalist picture. I shall not attempt to properly fix it now, but simply point out that personal computers similarly interact with their environment continually, but we still model them, as a whole and the individual programs they run, as Turing machines, which is due to the theory of *universal* Turing machines.

4 Arguments against computationalist claims

Claim 1: brain processes are computational.

This idea underlies most of cognitive science research that aims to replicate these 'brain computations', and is the focus of John Searle in [Sea90]. His stance can be summarised in this quote: 'The point is not that the claim "The brain is a digital computer" is false. Rather it does not get up to the level of falsehood. It does not have a clear sense' [Sea90, p. 35].

Computation can be defined as taking information (e.g. in the form of strings of 0s and 1s) and then processing and transforming it. A Turing machine, as a mathematical abstraction, does just that. But to say that an object in the physical world performs computation requires that certain aspects of it, e.g. voltages in wires, are treated as representing this information; Searle calls this *assigning syntax.* But this syntax is 'not intrinsic to physics' [Sea90, p. 27], but rather it is ascribed to a system by an observer. Furthermore, then, since this ascription of syntax is *prima facie* unrestricted, everything can, in principle, be interpreted as performing computation.

Unlike saying that the heart acts as a pump, which is a falsifiable proposition (under some reasonable interpretation of the word 'pump'), to say that the brain performs computations, then, is entirely empty, since it is trivially true under some assignment of syntax. The crux is that syntax, and hence the classification of something as a computer, are inherently observer-dependant.

A seeming way out of this is to assume that each human himself is the observer who uses their brain to compute with. The flow of information, as interpreted by this observer, is then from the outside world to the conscious perceptions

[1] This argument and the one above were made e.g. by psychologist Robert Epstein in aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer

of the observer. Indeed, most (if not all) models of e.g. vision work this way: the activations of retinal cells are quantified and algorithmically transformed to yield an output that should correspond to some conscious perception (e.g. see [CVR19]). But this presupposes a *homunculus* at some level: if the visual system 'outputs' information, in some syntactic form, it must be read by some homunculus that then actually *has* the conscious perception. Even if these perceptions are taken to be emergent from the 'computation', the problem remains that at some lower level, eventually, a homunculus is required to interpret states of the outside world as syntax that is to be processed.

I want to stress that the point here is not the classic mind-body problem, asking how physical states can lead to conscious perceptions (which we shall leave aside for now), but rather that in wanting to treat the brain as a computer, we are introducing an intermediate layer of syntax which physical states are to be interpreted as and in turn is to be perceived – when syntax itself is something that the observer assigns to the outside world. For personal computers, this circularity does not exist, as each (competent) user implicitly assigns syntax to the computer's physical states and thereby can make use of the computer. We would be putting the cart in front of the horse, however, if we were to say that perception (or cognition) works by performing computations on syntax, since this presupposes a user of the brain, which is, in Searle's words, to commit the homunculus fallacy.

Claim 2: the mind is software.

Given the brain-as-computer metaphor, it seems natural to think of the mind as the software running on this computer. Firstly, it is instructive to point out that we think of software not just as some sequence of computer instructions, but the outcome of executing these instructions on a computer; hence, simply pointing out that our conscious minds are clearly not just a sequence of instructions is not sufficient. Secondly, in talking of instructions that are executed, we are again running into the problem of assigning syntax; we shall leave this issue aside for now and consider the problems of the mind-as-software metaphor under the assumption that the brain can, in some sense, be thought of as an information-processing computer on which the mind is 'running'.

The first objection to this conception we shall consider is Searle's Chinese Room Argument [Sea80]: a person (who speaks no Chinese) is in a room and receives notes in Chinese; he has access to a book that gives instructions on how he should respond (in writing) to all kinds of notes, so that someone outside the room, who does speak Chinese and 'communicates' with the person inside the room by sliding notes under the door, would be under the impression that the person inside does, in fact, speak Chinese. Searle likens this to the working of a computer: the mere execution of instructions and syntax manipulation, as done by a computer or by the person in the room, does not by itself, he says, imply that there is *understanding*.

We might say that the person in the room is competent, but has no direct conscious experience (*qualia*) of this competence, as this competence does not stem from within him, but from (his interaction with) the book. Understanding I would define as the quale of competence, or competence with associated qualia. (Searle speaks of 'having semantics'.) It is not obvious what it would take for a computer, if at all possible, to have understanding/semantics, but it is clear, I belief in agreement with Searle, that following rules of a certain domain does not by itself imply understanding of that domain.

As mentioned at the outset of this section, we usually think of software as the outcome of the execution of a program. Here again the last section is relevant: this outcome is ultimately what we as the observer see. If the mind is software, where is the homunculus who observes it? It might be said that conscious experience is an *epiphenomenon* of the processes of the brain. What is incoherent to me about this view is that it cannot explain how we can have thoughts (as brain processes) about consciousness (this epiphenomenological realm) in the first place.

A popular proponent of epiphenomenalism is Douglas Hofstadter. He beliefs that computers might, eventually, be able to have minds similar to our own [Hof85]. To him, the essential building blocks of mind are symbols – as opposed to tokens (similar to Searle's semantics–syntax distinction). In a limited sense, he agrees with Searle's Chinese Room argument, in that mere token-manipulating programs do not have understanding. He does belief, however, that this token manipulation can lead to the creation of meaningful symbols in an emergent/epiphenomenological fashion.

Hofstadter takes issue with the mind-as-software metaphor in a different way, namely with the top-down approach it implies (and that is often assumed in cognitive modelling projects), where '[symbols] are manipulated by some overlaying program' [Hof85, p. 646]. To him, symbols emerge from the activities/manipulations of underlying tokens, and are themselves active entities: we (/our minds) are not agents that manipulate our symbols, but we *are* interactions of symbols (cf. [Hof07]). Further, he questions whether on the level of these symbols any computational rules can be said to be present; it might be that the activities that give rise to the symbols are *computationally irreducible* (a term I borrow from the very relevant [Wol02]). Hence, any modelling endeavour that uses abstractions at the level of those symbols, which are then 'shoved around', in Hofstadter's words, by some program, would fail to explain cognition.

In summary, Hofstadter beliefs that 'cognition is an activity that can be supported by computational hardware' [Hof85, p. 648], but rejects the idea of the mind as a symbol-manipulating program. Personally, I agree with his emergentist explanation of cognition, but find it unsatisfactory that Hofstadter explicitly equates 'having symbols' with consciousness, and sees thought experiments like Chalmers's p-zombies[2] as incoherent/non-issues (cf. [Hof07, ch. 22]).

[2] A *philosophical zombie (p-zombie)* is a thought experiment proposed by David Chalmers; a p-zombie is a being that is just like a normal person except that it has no *conscious* experience.

Claim 3: the brain can be modelled as a Turing machine.

This third issue is almost orthogonal to the previous two. We saw that Hofstadter beliefs that computers (which can clearly be modelled as Turing machines) could eventually have minds like our own, and Searle answers his third question in the affirmative.[3] A very insightful discussion on this issue can be found in Gödel's Gibb's lecture [GF86, pp. 304–323]. In it, Gödel talks about the philosophical implications of his incompleteness theorems. Specifically, he concludes that 'either [...] the human mind [...] infinitely surpasses the power of any finite machine, or else there exist absolutely unsolvable diophantine problems' [GF86, p. 310], and himself seems to favour the second disjunct.

To see where the 'diophantine' problems come from, should one assume, in accordance with the second disjunct, that the human mind is not more powerful than a Turing machine, it suffices to consider the Halting problem, i.e. the question of whether a given program halts on a given input. It is a result due to Turing that there can be no Turing machine that correctly decides all instances of the halting problem. In particular, if the human mind is at most as powerful as a Turing machine, it follows that there exists an instance of the halting problem that a human cannot correctly decide. (Allowing finitely many Turing machines to work together or to give them additional finite resources does not change this fact, so this result can be extended to humanity as a whole.)

A closely related problem is the (equally uncomputable) Busy Beaver function, that maps a natural number n to the number of consecutive 1s that a Turing machine (under a fixed definition) with n states could write on its output tape before eventually halting. Bringsjord [B$^+$06] (in an attempt to formalise what he calls 'Gödel's intuition', based on personal correspondence of Gödel and a very different reading of his Gibb's lecture that I cannot quite find evidence for) makes the case that humans are more powerful than Turing machines, because humans can, in principle, compute the Busy Beaver function for arbitrary values n. However, his argument rests on the explicit assumption that, if a human can compute the Busy Beaver function for n, they can do it for $n+1$.[4] This, in turn, rests on the assumption that humans can solve the halting problem for arbitrary machines ('Soon these [Turing machines of a particular size] will all be classified as halters or non-halters; for if need be [...] we can resort to manual machine-by-machine examination' [B$^+$06, p. 9]), hence committing *petitio principii* (which, interestingly, is also their rebuttal to one argument in their objection section). The reason I mention this paper is that, if not a formal proof, a semi-coherent argument can be made for claiming that human minds exceed the confines of Turing computability.

[3] Interestingly, Searle bases this intuition on the Church–Turing thesis ('anything that can be given a precise enough characterization as a set of steps can be simulated on a digital computer' [Sea90, p. 21]), but seems to suppose a stronger version that has been formulated by physicist David Deutsch, viz. that every physical process can be simulated by a Turing machine [Deu85].

[4] There seems to be an interesting parallel to the Sorites paradox here.

5 A tenable form of computationalism

With the above objections pointed out, is there any form of computationalism that is defensible? I belief there is, but it requires that we severely limit its claims in explaining (or even replicating) the mind. If we are to interpret the brain as a computer, we have to specify what exactly is to be interpreted as input and output – in Searle's words, we have to assign syntax – and that requires an observer, which cannot be the mind 'of that brain' itself. The way out, I think, is the way that is taken in all of experimental psychology: an experimental subject receives an input and produces a response as output, both of which can be interpreted by the experimenter, acting as the outside homunculus. With this recourse, however, we are abandoning our ambitions in explaining consciousness, an integral part of cognition – but there seems to me to be no alternative. Only in this way can we reasonably treat brains (at least others than our own) as information processing computers and investigate hypotheses about which functions they implement in whatever way.[5]

This evades the problems associated with both claims 1 and 2. As for claim 3, which I accept (encompassed under the stronger *Church–Turing–Deutsch thesis*, i.e. that every physical process can be simulated by a Turing machine [Deu85]),[6] I would claim that it seems very reasonable to accept Gödel's second disjunct, i.e. that there exist absolutely undecidable problems. These problems would be (or consist of) unknowable truths, and that these truths exist is also a (possible) conclusion to be drawn from Fitch's paradox (see e.g. [BS19]); other works, e.g. [Coo06], give credence to the existence of unknowable truths as well.

In summary, I think that we can understand the brain as a computer only *from the outside*. This suffices to investigate what functions the brain, understood this way, might implement to achieve its (human's) functioning in the world. However, concerning consciousness and the processes that are necessary and sufficient causes for it, a computationalist view will not get us far. Simply abstracting cognitive processes at convenient levels and simulating them cannot be enough to fully understand (conscious) cognition, let alone replicate it; both Searle and Hofstadter would, I think, agree. As far as this computationalist/cognitivist view has gotten us, and will likely continue to get us, in understanding the functioning of the brain in many different domains, a true science of the mind would be well-advised to find another approach.

[5] This seems awfully solipsistic: in treating subjects as black-boxes – physical input-output systems – we are completely disregarding their conscious experience. However, since it is impossible to know *for certain* that another person has conscious experience (i.e. other people are indistinguishable from p-zombies), I belief this to be justified.

[6] Of course, in labelling the processes of the brain as physical, I am myself evading the problem of how consciousness comes into the picture. I would like to address this problem in a future text.

References

[B$^+$06] Selmer Bringsjord et al. A new Gödelian argument for hypercomputing minds based on the busy beaver problem. *Applied Mathematics and Computation*, 176(2):516–530, 2006.

[BS19] Berit Brogaard and Joe Salerno. Fitch's Paradox of Knowability. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2019 edition, 2019.

[Coo06] Roy T Cook. Knights, knaves and unknowable truths. *Analysis*, 66(1):10–16, 2006.

[CVR19] Antonino Casile, Jonathan D Victor, and Michele Rucci. Contrast sensitivity reveals an oculomotor strategy for temporally encoding space. *ELife*, 8:e40924, 2019.

[Deu85] David Deutsch. Quantum theory, the church–turing principle and the universal quantum computer. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 400(1818):97–117, 1985.

[GF86] Kurt Gödel and Solomon Feferman. *Kurt Gödel: Collected Works: Volume III: Unpublished Essays and Lectures*, volume 3. Oxford University Press on Demand, 1986.

[Hof85] Douglas R Hofstadter. Waking up from the boolean dream, or, subcognition as computation. *Metamagical themas: Questing for the essence of mind and pattern*, pages 631–665, 1985.

[Hof07] Douglas R Hofstadter. *I am a strange loop*. Basic books, 2007.

[Kah11] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[Sea80] John R Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424, 1980.

[Sea90] John R Searle. Is the brain a digital computer? In *Proceedings and addresses of the American Philosophical Association*, volume 64:3, pages 21–37. JSTOR, 1990.

[SvL12] Keith Stenning and Michiel van Lambalgen. *Human reasoning and cognitive science*. MIT Press, 2012.

[vR08] Iris van Rooij. The tractable cognition thesis. *Cognitive science*, 32(6):939–984, 2008.

[Was68] Peter C Wason. Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3):273–281, 1968.

[Wol02] Stephen Wolfram. *A new kind of science*. Wolfram Media, 2002.