Mechanistic and computational explanations in neuroscience

the distinctness of computational explanation*

Anna Kocsis, mag.phil.
research assistant
Institute of Philosophy
Zagreb, Croatia

In their groundbreaking article „Computational Neuroscience“ Sejnowski, Koch and Churchland (1988:1300) are stating the following claim concerning the difference between mechanistic and computational explanations: „The chief difference is that a computational explanation refers to the information content of the physical signals and how they are used to accomplish a task.” In this talk I will argue for the distinctness of computational explanation in line with Sejnovski et al. and Chirimuuta's (2014) recent article in which he makes a case for the claim that computational explanations have explanatory power in virtue of them being *efficient coding explanations*. The plan of the talk is the following: first, I will present a particular take on the computation – information relationship using the analogy between Turing's machines and Sober's toy (1984:99). Then I will introduce the concept of *canonical neural computations* (CNCs), especially recent findings concerning one particular type of CNCs, namely *normalization*. Finally, I will argue for the distinctness of computational explanations based on information theory, specifically efficient coding.

As Chirimuuta (2014) notices, the idea that the brain is a computational device processing information is the most important dogma in neuroscience. It is frequently stated that computation *is* information processing, but, according to Piccinini and Scarantino (2010), it is important to keep these notions apart in order to understand the fundamental claims behind the nature of brain viewed as a computational *and* information processing device. Computation is an algorithmic operation which means that it works by following effective procedure. Expressed more informally, an algorithm is a collection of instructions or a recipe for carrying out some task. The notion of computation originated from mathematics, particularly from the development of the idea of a Turing machine (an abstract device that manipulates symbols on a strip of tape according to a table of rules). However, the idea of information processing originated within control and communication

* This work has been fully supported by the Croatian Science Foundation under the project number 5343.

engineering. Using the analogy between Sober's toy and Turing machines, I will present a particular perspective on the relationship between computation and information processing, one that will facilitate the usage of efficient coding explanations in establishing the distinctness of computational explanations in neuroscience.

To be able to characterize a Turing machine (TM), one has to be able to define and „make sense“ of its relevant states and transitions functions. Concerning the aforementioned analogy with Sober's toy, let's compare the screens of the game with the transition functions of a TM and the positions of the balls with the states of the TM. This analogy is a take on Johnson's (2010) view of computation where he calls the screens „information sinks“. If we want our characterization of tms to be of use in research about the nature of the brain and mind, a crucial distinction between tms and biological systems in general and brains in particular is worth keeping in mind. Brains were not designed ex novo with functions „written“ **into** them. Rather, they are a result of the system's operation as a turing machine under specific informational and structural circumstances over a long period of time. It is in understanding this constraint of the computer-brain comparison that the analogy with Sober's toy comes in handy.

Let's imagine we collect all balls, mix them together and drop them back into the toy. How can we go about characterising the process that takes place in the toy? If we would want to say something about the screens, we would need to know something about the balls first. Namely, we would characterise each sink by referring to the balls it keeps from passing through or the ones it allows to pass. If we would want to characterise the balls as parts of the process taking place inside the toy, we would have to know something about the screens. We would arrive to the relevant feature while describing balls (the size of the balls, not their colour) by studying the screens. As illustrated by this example, in biological systems, states and functions are intertwined. What is a relevant brain state or TM state is decided by the function in the sense that if you have a signal, an input, and there is no reaction (no function is activated) then the signal has no effect on the system and it is not a relevant state (that signal carries no informational content for the system). It follows that, on this view, information is about something but it is not necessarily representing something specific. Information has no semantic character in the sense that in this context it makes no sense to talk about misrepresentation or truth. Information is about the statistical character of a whole ensemble of messages, about the uncertainty in those messages and, as Jonhson (2010:653) points out, this is where „information and uncertainty find themselves partners“.

Why is this specific understanding of the computation-information interdependence helpful in understanding processes that take place in biological systems that compute? The reason is that this approach allows you to i) say why and in what sense a certain function (or transition function) is a „good“ response to the input and to make sense of the developed function, ii) (enables us to) assess the reasons behind the system selecting some states as relevant as opposed to others. Less formally, this treatment of information being processed is the „glue“ that bounds states and functions of a computing machine and allows formal description of the system. I maintain that this type of understanding is at the core of the argument for the distinctness of computational explanations based on efficient coding.

A specific situation in which this type of argument can be seen at work is the case of explanatory power of canonical neural computations (CNCs). Canonical neural computations are „standard computational modules that apply the same fundamental operations in a variety of contexts” (Carandini and Heeger, 2012:51) and serve as proof of concept for the thesis of computational modularity. Among many types of CNCs, some of the most familiar are exponentiation (thresholding), linear filtering and recurrent amplification. However, the one that has recently been the center of the attention in the scientific community is normalization (or divisive normalization). „Normalization computes a ratio between the response of an individual neuron and the summed activity of a pool of neurons” (Carandini and Heeger, 2012:51) where the normalized activity of each neuron is expressed by *the normalization equation*. Normalization was proposed by Heeger in the early 1990s to explain non-linear properties of neurons in the primary visual cortex. Normalization nas been assumed to be at work at the light adaptation process in the retina (Normann and Perlman,1979), size variance in the fly visual system (Reichardt et al., 1983), associative memory in the hippocampus (McNaughton and Morris, 1987), in the invertebrate olfactory system (Olsen et al., 2010) and is also argued to be a key feature of visual attention (Reynolds and Heeger, 2009) and context-dependent decision making (Louie et al., 2013).

As is emphasised by Chirimuuta (2014:138), due to accumulating evidence, it has become possible, to argue that „normalization is implemented by numerous biophysical mechanisms, depending on the system in question“ (for example, synaptic suppression and shunting inhibition are both possible implementations). In other words, normalization is multiply realized and the process can be described by a transformation of the normalization equation in each of the implementationally different instantiations. What is more interesting to observe is that each one of these situations results in some behavioural consequence (when I say behavioural, I mean a

computational end-point or output which can be a starting point of feedback information to the system in question). However, presumably we could get the same output without performing normalization. Differently put, there is no reason why the neurons that do not receive previously normalized input could not in principle develop functions that would generate the same output as in the case of the neurons that do receive previously normalized inputs. In line with Marr's (1982) views, there is a dissociation both between computation and implementation and computation and behavioral characteristics of the system. It seems that there are many possible computational solutions to a task or problem and many possible implementations of each of those solutions.

Recently, Kaplan (2011), Piccinini and Craver (2011) have argued against the distinctness of computational explanations claiming that the explanative strength of, for example, the normalization equation lies in model-to-mechanism-mapping. According to this view a computational explanation, namely the normalization equation, is explanative in virtue of the variables mentioned corresponding to identifiable components, organizational features etc. of the underlying implementational level. In the same vein, a computational expression is viewed as a mathematization of the underlying causal-mechanistic nature of the phenomena. I argue that, as is demonstrated by the case of normalization, the computational expression (the normalization equation) is not correctly viewed as an explanans and that the correct question concerning normalization should be „why normalize in the first place". The normalization equation is a mathematical expression obtained by a great number of measurements and data-fitting. Is has great predictive power – it will exactly predict the response of a neuron when the responses of local neurons are known – but has no explanative power. It is a formal description of the behaviour of the system. It makes no sense to map the variables in the description with causal elements of the process at the implementational level or to argue, based on the equation, for certain organizational features of the system. These mathematical regularities in the system are the very fact that needs to be explained. I argue that computational neuroscience and information theory do provide answers to the questions similar to the one presented concerning normalization and they do that using the notion of efficient coding.

As Johnson (2010) shows, the communication channel between neurons that are sensitive to different inputs gains capacity when those neurons are laterally connected, which normalization is an instance of. The precise mechanism of that connection is irrelevant, it can be achieved in different ways in different systems. What is important is that this gain in capacity allows the system to code information more efficiently and that is of seminal importance for a biological system heavily constrained with energy, size and time. This finding has a clear predictive force: each time a group of

neurons that are sensitive to different outputs is sending information to another neuron (or a different group of neutrons) they are likely to be laterally connected. This principle is instantiated by normalization. As a consequence, it is arguable that the system performs normalization on a certain group of neurons in order to gain channel capacity. This is a type of efficient coding explanation that explains the mathematical regularities observed in the system and formally expressed and quantified.

Primary literature:

Carandini, M. and **Heeger**, D. J. (2012) „Normalization as a Canonical Neural Computation" *Nature Reviews Neuroscience*, 13, 51–62.

Chirimuuta, M. (2014) „Minimal models and canonical neural computations: the distinctness of computational explanation in Neuroscience", *Synthese* 191:127–153.

Johnson, D. H. (2010) „Information Theory and Neural Information Processing", IEEE Transactions on Information Theory 56, No. 2: 653-666.

Kaplan, D.M. (2011) „Explanation and description in computational neuroscience", *Synthese* 183: 339–373.

Marr, D. (1982) *Vision: A computational investigation into the human representation and processing of visual information*, San Francisco: W.H. Freeman & Co. Ltd.

Piccinini, G., and **Craver**, C. (2011) „Integrating psychology and neuroscience: Functional analyses as mechanism sketches", *Synthese* 183(3), 283–311.

Piccinini, G. and **Scarantino**, A. (2010) „Computation vs. information processing: why their difference matters to cognitive science", *Studies in History and Philosophy of Science* 41: 237-246.

Sejnowski, T. J., **Koch**, C. and **Churchland**, P. S. (1988) "Computational Neuroscience", *Science* 241:1299–1306.

Secondary literature:

Louie, K., **Khaw**, M. W. and **Glimcher**, P. W. (2013) „Normalization is a general neural mechanism for context-dependent decision making" *PNAS* 110(15), 6139–6144.

McNaughton, B.L. and **Morris**, R.G.M. (1987) „Hippocampal synaptic enhancement and information storage within a distributed memory system" *Trends in Neuroscience* 10, 408–415.

Normann, R.A. and **Perlman**, I. (1979) „The effects of background illumination on the photoresponses of red and green cones" *Journal of Physiology* 286, 491–507.

Olsen, S.R., **Bhandawat**, V. and **Wilson**, R.I. (2010) „Divisive normalization in olfactory population codes“ *Neuron* 66, 287–299.

Reichardt, W., **Poggio**, T. and **Hausen**, K. (1983) „Figure-ground discrimination by relative movement in the visual system of the fly. Part II. Towards the neural circuitry“ *Biological Cybernetics* 46, 1–30.

Reynolds, J.H. and **Heeger**, D.J. (2009) „The normalization model of attention“ *Neuron* 61, 168–185.

Sober, E. (1984) *The Nature of Selection: Evolutionary Theory in Philosophical Focus*, Bradford/MIT Press.