The brain as a model of the world

Oron Shagrir[1]

Abstract. My aim here is to show that an underlying assumption in computational approaches in cognitive and brain sciences is that the brain is a model of the world in the sense that it mirrors certain mathematical relations in the surrounding environment. I will give here three examples. One is from David Marr's computational-level theory of edge-detection. The second one is the computational work on the oculomotor system. And the third one is a Bayesian model of causal reasoning. One might wonder why this brain-as-a-model-of-the-world assumption is so prevalent in computational cognitive science and neuroscience. My proposed answer (for which I will not argue here) is that in these fields computation just means a dynamical process that models another domain. Thus saying that the brain computes just means that its processes models certain mathematical, or other high-order, relations in another domain, often the surrounding world.

1 MARR'S COMPUTATIONAL LEVEL

In *Vision*, Marr famously advances a three-level approach to the study of visual processes. The computational level specifies *what* is being computed and *why*. The algorithmic level characterizes the system of representations that is being used, e.g., decimal vs. binary, and the algorithm for the transformation from input to output. The implementation level specifies how the representations and algorithm are physically realized. What Marr meant by these levels and how he saw the relations between them have been a topic of debate for many years. My focus here is the most distinctive, and least well understood, of Marr’s three types of analysis, which is the computational level. Marr's notion of computational-level theory has received a variety of interpretations. Many have argued that the computational level aims at stating the cognitive phenomenon to be explained; the explanation itself is then provided at the algorithmic and implementation levels [1, 2, 3]. Others have described it as providing sketches of mechanism [4]. Yet others have associated the computational level with an idealized *competence* and the algorithmic and implementation levels with actual performance [5, 6]. Finally, Egan [7] associates the computational level with an explanatory formal theory, which mainly specifies the computed mathematical function. I have defended a different interpretation that emphasizes the role of the environment in Marr's notion of computational analysis [8,9]. Marr characterizes the computational type of analysis as specifying “what the device does and why” ([10], p. 22) Most commentators have addressed the *what* aspect, and I agree with

Egan that Marr aims to characterize the mathematical function that is being computed. According to the computational theory of edge-detection, for example, early visual processes compute the zero-crossings of (Laplacian) second derivative filterization of $\nabla^2 G*I$ (∇ is a Laplacian, G is a Gaussian and I is the retinal image). Marr, however, repeatedly insists that computational-level theories also include the *why* aspect whose aim is to demonstrate the basis of the computed function in the physical world ([11], p. 37). Marr associates this *why* aspect with what he calls *physical constraints*, which are physical facts and features in the physical *environment* of the perceiving individual ([10], pp. 22-23). These are constraints in the sense that they limit the range of functions that the system could compute to perform a given visual task successfully.

What exactly are the relations between the physical constraints and the computed function? How do these constraints substantiate the basis of the computed function in the physical world? The gist of my interpretation is that Marr's working hypothesis is that the visual system is a model of the world in the sense that it mirrors or preserves certain structural relations in the visual field. By *structural relations* we mean "high order" mathematical, geometrical, or other formal relations. The visual system would *preserve* these relations if there were an isomorphic mapping from the visual system onto the visual field; more realistically we talk about homomorphism or partial-isomorphism and even these mappings involve a vast amount of approximation and idealization. Our claim is that a computational analysis appeals to the physical constraints in order to underscore these morphism relations, and in doing this, these constraints play both explanatory and methodological roles in theories of vision. Explanatorily they serve to demonstrate the appropriateness and adequacy of the computed function to the information-processing task ([10], pp. 24-25). Methodologically they serve to guide discovery of the function that the visual system computes [12].

Thus to take edge-detection, the mathematical function being computed (zero-crossings of second-derivatives) reflects sharp changes in light reflection in the visual field that often occur along physical edges such as object boundaries (whereas the latter changes can be described in terms of extreme points of first-derivatives or zero-crossings of second derivatives of the reflection function; see figure 1).

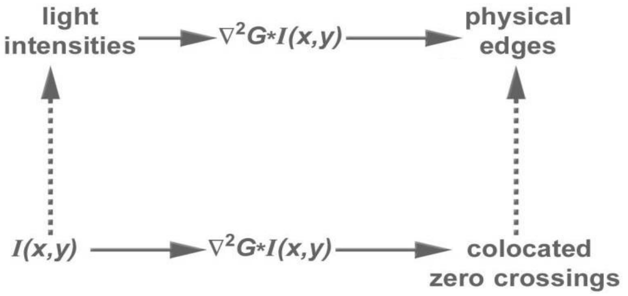

Figure 1. **Edge-detection.** Early visual system detects edges by computing the zero-crossings of the function $\nabla^2 G*I$ (lower-span), whereas different filters come with different Gaussians, G. The elements ("pixels") that constitute retinal image, I(x,y), encode (dashed arrow) light intensities in the visual field. The visual edges (formed by segments of zero-crossings) encode physical edges such as object boundaries. The relation between light intensities and physical edges can be described too in terms of extreme points of first derivatives or zero-crossings of second derivatives (lower-span), as illumination and reflection often change sharply along (say) object boundaries.

[1] Dept. of Philosophy and Cognitive Science The Hebrew Univ., 91905, Jerusalem, Israel. Email: shagrir@cc.huji.ac.il.

This physical fact ("constraint") – that sharp changes in reflection often occurs along physical edges – explains why the visual system computes derivation, and not (say) factorization or exponentiation, for the task of edge-detection. It also guides the visual theorist in discovering the mathematical function that the system computes, namely, derivation.

Marr never discusses isomorphism or structural similarities explicitly. Nevertheless, I have shown that it is central to his computational analysis of other visual tasks as well. My aim here is to show that Marr's notion of computational analysis is not confined to vision but is widely applicable in computational brain and cognitive science.

2 THE NEURAL INTEGRATOR IN THE OCULOMOTOR SYSTEM

The neural integrator converts eye-velocity inputs to eye-position-outputs, and thus enables the oculomotor system to move the eyes to the right position [13, 14]. The inputs arrive from fibers coding vestibular, saccadic or pursuit movements.

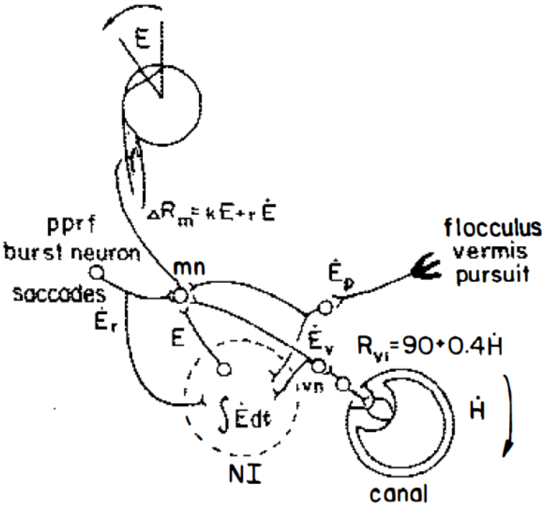

Figure 2: The neural integrator (NI). NI receives eye-velocity coded inputs, $\dot{E}$, and, computing integration, produces eye-velocity coded outputs, E. The hypothesis is that the integrator is common to vestibular, saccadic and pursuit movements, thus receiving vestibular ($\dot{E}_v$), saccadic ($\dot{E}_r$), and pursuit ($\dot{E}_p$), velocity coded inputs. On the right it is shown how the head velocity signals, $\dot{H}$ are converted into eye-velocity codes ($\dot{E}_v$). These codes are projected directly to the motoneurons (mn) that have to produce velocity commands, but also the neural integrator (NI) which produces position codes projected to the motoneurons for position commands (from [13], p. 35).

The system produces eye-position codes by computing mathematical integration over these eye-velocity encoded inputs. Take vestibular movements, where the eyes are moved in the same velocity and opposite direction to head movements. The task of the integrator is to compute the new eye-position based on the velocity signals transduced through the canals behind our ears. On the right side of figure 2 it is shown how the head velocity signals, $\dot{H}$ are converted into eye-velocity codes ($\dot{E}_v$). These codes are projected directly to the motoneurons (mn) that have to produce velocity commands, but also the neural integrator (NI) which produces position codes projected to the motoneurons for position commands. In cats, monkeys, and goldfish, the network that computes *horizontal* eye movements appears to be localized in two brainstem nuclei, the nucleus prepositushypoglossi (NPH) and the medial vestibular nucleus (MVN).

Mathematical integration characterizes operations performed in two *very different* places. One is in the neural representing system, namely, the neural integrator. It performs integration on the neural inputs to generate neural commands. This is of course the reason that the system is known as *integrator*. Another and very different place, however, is in the target domain being represented, in our case the eyes. The relation between position and velocity of the eye can be described in terms of integration too! The distance between the previous and current positions of the eye is determined by integrating over its velocity with respect to time. So what we have here is an (iso-)morphism between the representing sensory-motor neural system (the integrator) and the represented target domain (the eyes and their properties). The neural integrator mirrors or preserves certain relation in the target domain, namely the distances between two successive eye positions. By computing integration, the neural function mirrors, reflects or preserves the integration relation between eye velocity and eye positions.

Let us put these findings in the context of Marr's notion of computational analysis. The *what* aspect describes the mathematical function, integration, computed by the neural integrator. The *why* aspect relates the computed function with the physical environment, namely, the eyes with their properties. The analysis invokes a physical constraint, which in our case is the velocity-position relation, namely the relation between eye-velocity and the distance between two successive eye-positions. Using this constraint, it is shown that there is a morphism mapping between the neural function and the target domain. This mapping relation is underscored by the fact that the two domains have a shared structure, which is mathematical integration.

As said, the physical, environmental, constraints play both explanatory and methodological roles. On the explanatory side, they serve to explain *why* computing integration is appropriate for the task of controlling eye movement. The neural network computes integration and not, say, multiplication, exponentiation, or factorization, *because* integration preserves the velocity-position relation, namely, the integration relation between eye movement and eye positions in the target domain. Factorizing numbers would not result in moving the eyes to the right place, precisely because it does not preserve relations in the target domain that are relevant to eye movements. Integration does: When you compute integration over eye-velocity encoded inputs, you mirror the integration relation between velocity and position; hence, you output representations of a new eye position. The algorithmic and implementation levels complement this explanation by specifying how this integration function is carried out in the neural system.

On the methodological side, the velocity-position relation is instrumental in discovering what function is computed. In our example, experimental electrophysiological results indicated that the neural system converts eye-velocity pulses into eye-position codes. Looking at the relation between the represented velocity and position, theoreticians quickly inferred that the internal relations between the representing states must be of integration. This logic of discovery assumes that the computed function is that of integration since the computed function must correspond to the velocity-position integration relation, which is already known.

3 A BAYESIAN MODEL OF CAUSAL REASONING

Bayesian models of cognition have taken a central place in cognitive science. Their aim is to explain how humans should update their beliefs in the face of sparse experiential data. The models combine two trends. One is a rationality or optimality analysis that aims to single out the optimal solution for a given problem. In the context of cognition the aim is to account for how a rational agent *should* reason in a situation of uncertainty. The assumption is that this normative account will also tell us something about how the agent does reason, for example, by approximating the optimal solution. The other trend is a probabilistic approach that aims to explain inferences in terms of subjective probabilities, namely, degrees of belief. The idea, in general, is to update the probabilities ("degrees of belief") in some hypotheses space with regard to incoming data about the world; the central tool in calculating these probabilities is usually the Bayes' rule. Bayesian models have been invoked to account for human reasoning, but also for many other aspects of cognition, including problems in concept formation, visual perception, motor control, language processing, causal learning and reasoning, and even social cognition.

To give you some flavor of it, assume that we want to model how an ideal doctor *should* reason given certain intuitive assumptions about causal relations between various variables, some of which are overt and some are hidden (fig. 3). In this example, the observed events are risks and symptoms, and the hidden ones are diseases. This particular model assumes certain intuitive causal principles that constraint the structure. The arrows in the structure represent hypotheses about the causal relations, that are assigns certain probabilities. The structure (in black) represents the prior assumptions and probabilities about this kind of causal reasoning. The model updates probabilities when more data is entering. In this example, the red arrows represent new hypotheses that are added to the structure given the new data about the patient who works in the factory and suffers from chest pain.

Interestingly, Bayesians often compare their approach to Marr's conceptual framework [15, 16]. One thing that they mean by this comparison is that they aim to specify a formal or mathematical function that is in some sense an optimal solution for the cognitive task at hand. In our example, the pertinent mathematical function is provided in terms of the directed graph. They also mean that, at this level, they do not aim to specify what kinds of representations are used, what algorithms are performed and how they are implemented in the brain. These issues belong to the algorithmic and implementation levels (which complements the computational level).

But I think that that the analogy between the Marrian and the Bayesian approaches extends to the *why* aspect in the Marr's computational analysis, namely, to the relations between the computed mathematical function is related to the environment. In their programmatic paper, Griffith, Kemp and Tenenbaum [17] write that the big computational question that underlies the Bayesian approach is "How does the mind build rich, abstract, veridical **models of the world** (my emphasis) given only the sparse and noisy data that we observe through our senses?" Indeed, looking again at the model for causal reasoning we should note that models implicitly refers to two different domains. Firstly, the model represents a cognitive domain: The assumption is that a cognitive agent should employ a scheme that has the structure of a directed graph (again, how the structure is implemented in psychological and neural mechanisms is not part of the computational theory). In other words, the model specifies the parameters that the cognitive domain must employ and the functional relations between them (that are presented in terms of a directed graph). But, now, the cognitive domain itself is a representational system: It represents the world, in our case classes of risks, diseases and symptoms, and the causal relations between them.

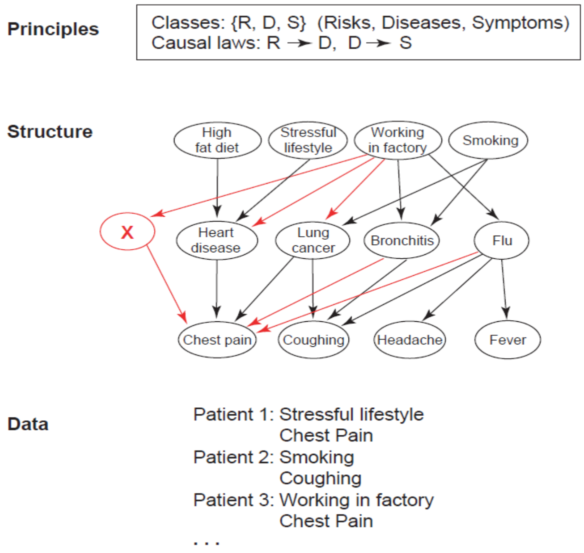

Figure 3: Bayesian causal induction. Knowledge in a medical domain is represented using directed graph. There are two classes of observed variables – risk factors, and symptoms – and one class of hidden variables – diseases – with causal relations from risks to diseases and diseases to symptoms only. Given a newly observed correlation (e.g. between working in a factory and chronic chest pain), the graph generates a constrained set of hypotheses for how that data might be explained (shown in red). The model represents an idealized cognitive inferential structure, which in turn, represents the causal relations between the three classes of variables (From [16], p. 315).

In the presentation of the model this distinction between the latter two domains, the cognitive system and the world, is blurred. In fact, it is ambiguous whether "Flu → Fever" refers to a certain relation in the cognitive domain or in the world. This ambiguity takes place not because the relations are the same (they are not: a representation of flu is not sick). The ambiguity is innocuous because the implicit assumption is that there is a sort of morphism between the cognitive domain and the world. The assumption is that the inference relations between a representation of flu and a representation of fever mirrors the causal relations between flu and fever. Both relations are, for example, non-symmetric: The relation holds in one direction, but not necessarily in the other direction. So we see that Bayesians too assume that the brain (in this case the cognitive system) is a model of the world; it is a model of the world in the sense that it preserves certain high-level formal relations in the environment.

CONCLUSION

Is computation observer relative? Before answering this we should clarify what is meant by computation. I have shown that the idea that the nervous system models the world is quite prevalent in different computational approaches in cognitive and brain sciences, whereas 'model' here means a representational system that preserves structures in the world. One might wonder why this brain-as-a-model-of-the-world assumption is so prevalent in computational cognitive science and neuroscience. Elsewhere I argue that at least in cognitive and brain sciences computation just *means* a dynamical process that models another domain. Thus saying that the brain computes means that its processes models certain mathematical, or other high-order, relations in another domain, often the surrounding world.

We can now return to the question about observer-relativity. In one sense computation is observer elative. An external observer could use, as it were, a physical system to model some other domain, mathematical or physical, that is isomorphic to it. Thus one could use the neural integrator, for example, to compute integration. In other senses computation is not observer relative. I can use the integrator to compute integration, but (arguably) I cannot use it to compute some other mathematical functions. Thus the function that is being computed is not observer relative. More importantly, there are cases in which the computation itself is not observer relative. That the neural integrator computes eye-positions from eye-velocity is not observer relative; it is a matter of objective fact about the brain. I hope to expand on these tentative remarks in the talk.

REFERENCES

[1] J. L. Bermúdez. *Philosophy of psychology: a contemporary introduction*. New York: Routledge (2005).

[2] S. W. Horst. The computational theory of mind. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*: http://plato.stanford.edu/entries/computational-mind/. (2009).

[3] W. Ramsey. *Representation reconsidered*. Cambridge, UK New York: Cambridge University Press (2007).

[4] G. Piccinini and C. Craver. Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese, 183*, 283-311 (2011).

[5] T. W. Polger. Neural machinery and realization. *Philosophy of Science, 71*, 997-1006 (2004).

[6] C. F. Craver. *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. New York: Oxford University Press (2007).

[7] F. Egan. Computational models: a modest role for content. *Studies in History and Philosophy of Science, 41*, 253-259 (2010).

[8] O. Shagrir. Marr on computational-level theories. *Philosophy of Science, 77*, 477-500 (2010).

[9] O. Shagrir and W. Bechtel. Marr's computational-level theories and delineating phenomena. In D. M. Kaplan (Ed.), *Integrating psychology and neuroscience: Prospects and problems*. Oxford: Oxford University Press (2014).

[10] D. C. Marr. *Vision: A computation investigation into the human representational system and processing of visual information*. San Francisco: Freeman (1982).

[11] D. C. Marr. Artificial Intelligence - personal view. *Artificial Intelligence, 9*, 37-48 (1977).

[12] E. C. Hildreth and S. Ullman. The computational study of vision. *Foundations of cognitive science* (pp. 581-630): MIT Press (1989).

[13] D. A. Robinson. Integrating with neurons. *Annual Review of Neuroscience, 12,* 33-45 (1989).

[14] R. J. Leigh and D. S. Zee. *The neurology of eye movements* (4th ed.). New York: Oxford University Press (2006).

[15] N. Chater, J. B. Tenenbaum and A. Yuille. Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences, 10*, 287-291 (2006).

[16] J. B. Tenenbaum, T. L. Griffiths and C. Kemp. Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science, 10*, 309-318 (2006).

[17] T. L. Griffiths, C. Kemp and J. B. Tenenbaum. Bayesian models of cognition. In R. Sun (Ed.), *Cambridge handbook of computational cognitive modeling* (pp. 59-100). Cambridge: Cambridge University Press (2008).