The Brain as an Input-Output Model of the World

Oron Shagrir

My aim 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 relations in the surrounding world. Elsewhere (Shagrir 2012) I argue that the notion of *internal model* is quite central in computational cognitive neuroscience. The focus here is on a weaker form of modelling, called input-output modelling, in which the input-output function of a nervous system mirrors a certain relation in the target domain. I argue that the input-output modelling assumption is entrenched in computational approaches, playing distinct methodological and explanatory roles in cognitive neuroscience. Methodologically, input-output modelling serves to discover the computed function from environmental cues. Explanatorily, input-output modelling serves to account for the appropriateness of the computed function to the explanandum information-processing task.

The paper proceeds as follows: I start with some general discussion about modelling (section 1). I next present three examples in which the theoreticians assume (or so I argue) input-output modelling (section 2). Next, I argue that input-output modelling plays a distinct theoretical role in cognitive neuroscience. It often plays a methodological role in discovering the internal computed function (section 3), and it even more often plays an explanatory role in accounting for the appropriateness of the computed function to the explanandum cognitive task (section 4). I finally compare the modelling approach with mechanistic explanations (section 5) and with interpretative models, or *I*-models (section 6).

At the outset, I would like to make clear what the paper is and is not about. When talking about modelling in cognitive neuroscience, it is crucial to distinguish between the claim that scientists use models to study, simulate and explain the brain and the claim that the brain itself is a model of the world. According to the first claim, the brain is the target domain and, as such, it is no different from many other physical and biological systems that scientists model. According to the second claim, the brain itself is a modelling system. In this respect the brain is distinctive: There are hardly any other natural systems, if any at all, that are

models of other domains. Although I will examine several models of the nervous system, the aim of the paper does *not* address the (first) claim about scientific models. The paper is about the second claim, namely, that the brain itself is a model of the world. The aim is to explicate this second-sense modelling relation, in which the nervous system is a model of the world. Obviously, the two claims are related in that computational models of the brain depict the brain as a model of the world (or so I argue). My focus, however, is the brain-as-model-of-the-world view. The "world" – the target domain – is to be understood in a very broad sense. It can be part of the immediate environment of the organism, e.g., the visual field. But the world can be more distant parts, as well as future, past, and even imaginative and counterfactual scenarios. The world can also be some internal bodily or mental states; in our main example, the target domain ("world") consists of properties of the eyes.

I should also clarify that the aim of the paper is to not to defend the brain-as-a-model view. I am not suggesting here that the view is empirically adequate, nor do I claim that the computational examples discussed here are impeccable. The aims are to say what the view means, to show its centrality in computational work in cognitive neuroscience, and to elucidate its distinct theoretical role in studying the nervous system.

1. Input-output modelling

There is a wealth of literature about models.[1] I will take a model to be a representational system that preserves patterns of relations in the target, represented, system. By preserving patterns of relations I mean that there is an isomorphism – or more realistically something less than that – from the representing system onto the target system. Another way to put this is to say that the model and the target domains share some structural – formal or mathematical – similarities (Swoyer 1991).

A nice example of a model is a family tree (fig. 1). In this tree the lines, arrows, and double arrows preserve certain familial relations like being a sibling, being a parent and being married. This does not mean, of course, that the relations of the models are exactly similar to the familial relations. Hopefully, being married is not exactly the same as being related by a double arrow. The similarity is at a higher, structural level, of a mathematical or formal similarity. In our case, being married and being related by a double arrow are both

[1] See, e.g., Frigg & Hartmann (2017); Weisberg (2013).

symmetrical relations. The family tree also exemplifies the main function of models, which is surrogative reasoning. What this means is that we use models to reason about the target domain; our inferences about the target are made by looking at the model, not at the target. We can infer whether John is or is not the grandparent of Mo by looking at the model alone. This is made possible precisely because the relations in the model preserve, or mirror, relations in the target domain.[2]

While most people would agree that some amount of structural similarity is *necessary* for modelling, they would also argue that the demand for a full-fledged isomorphism is excessive, at least when we talk about concrete models. Most people thus confine the modelling requirement to homomorphism, partial isomorphism, or even some weaker morphism relations.[3] I focus here on input-output modelling, which requires a minimal amount of morphism. We will say that a system is an input-output model just in case it satisfies the following conditions:

(1) There is a process of the system that transforms input variables to output variables.[4]
(2) The input-output function, *f*, preserves a certain relation, $\underline{R}$, in a target domain: There is a mapping from the model onto the target domain that maps *f* to $\underline{R}$, *x* to $\underline{x}$, *y* to $\underline{y}$,…, such that $f(x)=y$ iff $\langle \underline{x},\underline{y} \rangle \in \underline{R}$.
(3) The mapping is a representation relation in that the input and output variables, x and y, represent the features, $\underline{x}$ and $\underline{y}$, in the target domain.

My focus here is condition (2), namely, showing that an underlying assumption in computational work in cognitive neuroscience is that the nervous system satisfies (2). As for condition (1), it is certainly not true that all brain activity is described in terms of input-output processes; there are for instance endogenous neural mechanisms (Bechtel 2012). But it is not controversial that broadly speaking much of brain activity is couched in terms of input-output processes. I will simply focus here on the latter processes. As for (3), it is certainly controversial whether the nervous system "really" represents. Scientists describe

[2] See Swoyer (1991) for a general discussion about the relation between modelling and surrogative reasoning. See Grush (2004) for a discussion about modelling and surrogative reasoning in the brain.
[3] Less-than-isomorphism characterizations are in terms of partial isomorphism (French and Ladyman 1999; Da Costa and French 2003), homomorphism (Bartels 2006), and similarity (Giere 2004).
[4] The inputs and outputs need not be peripheral to the brain. In some examples discussed below we talk about sub-systems whose inputs are received and/or their outputs are projected to other parts of the nervous system. The inputs and outputs are very often (magnitude) values of certain properties such as voltages.

the nervous system as representing, carrying information about, or encoding values, properties and objects in the world. Often, these notions refer to some causal-based relations, and then the controversy is whether these causal-based relations amount to representations (e.g., Ramsey 2007). I will not get into this controversy. I will take it that (3) is satisfied when the scientists describe the nervous system as representing (which often points to some causally-based relation). My contention is that scientists often assume that the brain is *a certain kind* of representational system, namely, that it is a system that bears morphism relations to the target world (condition (2).

A few words about condition (2): Throughout the paper I will use the italicized symbols such as *x* and *y* to signify some properties of the representing system, and underlined italicized symbols such as $\underline{x}$ and $\underline{y}$, to signify properties in the target domain. The relations *f* and $\underline{R}$ in the condition signify some physical relations in the representing and target domain (respectively). The condition amounts to saying that there is a similarity here in the more abstract, e.g., formal, level. Some might say that at the more abstract level, *f* and $\underline{R}$ are similar formal relations; for example, that both are mathematical integration. One way or another, this similarity should be taken with a grain of salt. Given that we are talking here about domains – both the models and their targets – that are biological and physical systems, these similarity relations involve a vast amount of approximation and idealization.

Cummins (1989) presents a somewhat similar notion, of *input-output representation*, in his famous Tower-Bridge diagram (fig. 2).[5] Ramsey further introduces the notion of an internal model that is often associated with the term *structural representation* or *S-representation* (Swoyer 1991; Ramsey 2007). These authors, however, aim to account for the notion of representation, and specifically mental or cognitive representation. They make the further claim that morphism (or a sufficient amount of it), perhaps with some other conditions, is constitutive for being a representation.[6] A well-known argument against the sufficiency of isomorphism is that a system that is isomorphic to one target domain is immediately

[5] The term *input-output representation* is coined by Ramsey (2007: 68-77), who associates it with task analysis.

[6] See also Gallistel and King: "Representations are functioning homomorphisms. They require structure-preserving mappings (homomorphisms) from states of the world (the represented system) to symbols in the brain (the representing system). These mappings preserve aspects of the formal structure of the world" (2009: *x*).

isomorphic to many other target domains without representing or modelling them; relatedly, isomorphism is a symmetric relation whereas representing and modelling are not.[7]

As implied above, this paper is *not* in the business of analyzing the relations between morphism, representing and modelling. In particular, I do not argue that morphism is necessary and/or sufficient for representing. Nor do I argue for the satisfaction of condition (3) above, namely, that the brain is a representational system. I will thus refrain from using the terms *structural representation* and *input-output representation*, which are often accompanied by the philosophical baggage that morphism is necessary and/or sufficient for being representation. My claim is that much of the computational work in cognitive neuroscience assumes that the nervous system is morphic to the world, in the sense of condition (2). And given that this work also describes the brain as representing, carrying information about, or encoding, we can say that it assumes that the brain is a model of the world.

There are many philosophers and scientists who advance claims about the nervous or cognitive system modelling the world. Cummins (1989), Ramsey (2007) and Gallistel and King (2009) associate this modelling idea with more classical theories of cognition. Others note that the notion of a model is central in non-classical theories as well (e.g., Eliasmith and Anderson 2003; Grush 2004; Ryder 2004; Churchland 2007; O'Brien and Opie 2009; Shagrir 2012). Perhaps the best-known example of a model is the cognitive maps in the hippocampus of rats, humans and other mammals and animals. These maps, which consist of place cells, are used for navigation and spatial processing (O'Keefe and Nadel 1978). The idea that the brain (or mind) is a model of the world is also pivotal in Bayesian approaches to cognition.[8] Though it is questionable that they all use the same notion of a model, most of these scientist and philosophers pose *internal models*, in which some of the internal relations within the system also preserve relations in the target domain. Posing these models serves to explain certain cognitive phenomena such as motor control.

[7] See, e.g., Suárez (2010).

[8] Thus Griffith, Kemp and Tenenbaum (2008) say that the big computational question that underlies the Bayesian approach is "How does the mind build rich, abstract, veridical models of the world given only the sparse and noisy data that we observe through our senses?". See also Clark (2015), who further emphasizes the central role of generative models in the hypothesis that the brain is a prediction machine.

My claim here is that a far more widespread assumption is that the nervous system models the world in the weaker sense of input-output modelling. This assumption is found in many cases in which no internal models are posited. Moreover, these input-output morphism relations play a distinct theoretical – both explanatory and methodological – role in computational work in cognitive neuroscience. My aim here is to exemplify this role, which is not emphasized enough by theoreticians and is often ignored by philosophers.

2. Three examples

Fodor (1994) and others (Haugeland 1981; Pylyshyn 1984) stress the fact that a digital computer (or a Turing machine) has the ability to support processes that are truth-preserving. This means that we can implement in these systems inference ("syntactic") rules that mirror semantic relations such as logical validity. Taking the inputs and outputs to be symbolic expressions (say, the input is a set of sentences K and the output is a sentence *p*), the "inner" input-output function, which is the inferential relation from K to *p*, mirrors the extensional semantic relations; in other words, $K \vdash p$ iff $K \vDash p$:

> Well, as Turing famously pointed out, if you have a device whose operations are transformations of symbols, and whose state changes are driven by the syntactic properties of the symbols that it transforms, it is possible to arrange things so that, in a pretty striking variety of cases, the device reliably transforms true input symbols into output symbols that are also true (Fodor 1994: 9).

Cummins (1989) and Ramsey (2007) associate input-output modelling with classical theories of cognition. But as I will show below, the input-output modelling occurs also in "non-classical" theories where the content of the representations is often non-propositional, and so the input-output relations are not truth-preserving. Nevertheless the input-output relations preserve relations in the target system in the sense of morphism stated above.[9] In what follows, I describe three non-classical cases of input-output modelling in cognitive neuroscience.

The neural integrator in the oculomotor system. The oculomotor system controls eye movements. There are several types of eye movement. Gaze stabilization movements stabilize the visual world on the retina when the head/body is moving. The *vestibulo-ocular reflex* (VOR) keeps the visual world stable on the retina when the head is moving. The *opto-kinetic reflex* stabilizes the visual world when the head is stationary (e.g., when one is

[9] We can say that truth-preserving is just a special case of the morphism relation.

looking out from a train's window). Gaze-aligning movements include voluntary and reflexive saccades and smooth pursuit movements that allow one to track a moving target (Glimcher 1999; Leigh and Zee 2006). Our focus is a sub-network of the oculomotor system called the *neural integrator*. It receives as inputs neural signals that encode velocity and transform them to signals that encode position. The neural integrator converts eye-velocity inputs into eye-position outputs and thus enables the oculomotor system to move the eyes to the right position (Robinson 1989; Seung 1998; Eliasmith and Anderson 2003; Leigh and Zee 2006).

Take vestibular movements, where the eyes are moved in the same velocity as, and opposite direction to, head movements. A wealth of experimental evidence from the 1960s onward indicates that the vestibulo-ocular system determine the new eye position on the basis of inertial velocity information transduced through the canals behind our ears (the semicircular canals). 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).[10] Robinson and others infer that this velocity-to-position function is performed by an integrator network (I discuss the logic behind this inference in section 3). Thus Robinson writes:

> That there is indeed a second integrator is without doubt, since single unit studies in the vestibular and abducens nuclei show that the firing of units in the vestibular nuclei are in fact proportional to head velocity (over the bandwidth mentioned) and single units in the abducens nuclei increase their rate of firing in a manner proportional to eye position during the slow phase of nystagmus for which the lateral rectus is an agonist (1968: 1041)

Robinson (1989; Cannon and Robinson 1987) also hypothesizes that the same neural integrator is used for vestibular, optokinetic, saccadic and pursuit movements (fig. 3).[11] The inputs arrive from different fibers coding vestibular, optokinetic, saccadic and pursuit velocity. The integrator system produces eye-position codes by computing mathematical integration over these eye-velocity encoded inputs. On figure 3, the eye-velocity codes, $\dot{E}$, are projected directly to the motoneurons that have to produce velocity commands in order to move the eyes in the right speed. But the eye-velocity codes, $\dot{E}$, are also projected to the neural integrator that produces position codes, E. The latter eye-position codes are further projected to the motoneurons for position commands.

[10] See the reviews by Robinson (1968; 1989) and the one by Leigh and Zee (2006)
[11] See also Goldman et al. (2002).

Crucially, 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 the reason that the system is known as an *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 input-output modelling. The input-output function of the representing sensory-motor neural system (the integrator) mirrors or preserves a 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.[12]

We can describe this morphism relation between the representing neural system and the represented target domain (the eyes and their properties) in the framework of the Tower-bridge picture (fig. 4). The lower span describes a causal process in the neural system (i.e., in the neural integrator) that transforms input values, $\dot{E}$, that code eye velocity, $\underline{\dot{E}}$, to output values, E, that code eye position, $\underline{E}$.[13] The computed function is mathematical integration, namely, the values E, are the result of mathematical integration over $\dot{E}$ with respect to time. The upper span describes a certain relation in the target domain, namely, the eyes. The new position (which is the distance from the previous position), $\underline{E}$, is also a result of mathematical integration over the velocity, $\underline{\dot{E}}$, with respect to time. Thus the mapping relation, I, which maps the input values, $\dot{E}$, to the encoded velocity values, $\underline{\dot{E}}$, and the output values, E, to the encoded distance values, $\underline{E}$, is a morphism relation.

Path integration. Our neural integrator is by no means unique. Our brain computes mathematical integration to solve other problems as well. Homing is the ability of animals and humans to return to their departure point. Animals use external cues – environmental

[12] To keep things simpler, I will use here the terms *distance* and *position* interchangeably. New (horizontal) position is evaluated on the basis of the distance from the previous position.

[13] Note that in figure 3 the term E stands for both the representing (output) neural activity and the represented eye position. Similarly the term $\dot{E}$ stands for both the representing (input) neural activity and the represented eye velocity. This presentation is customary in neuroscience. This sort of presentation underscores (again) the modelling assumption, as it is apparent that the integration relation holds in both representing and represented domains.

stimuli and events – to navigate back home.[14] But experimental results show that homing occurs even when all the external cues are removed. Cues about initial reference and self-motion suffice to calculate the animal's relative spatial location; this phenomenon is called *path integration*.[15] The input of the calculation is angular velocity signals, which are provided by the vestibular or other systems. In this case, there might not be a specific neural sub-system that computes integration. Nevertheless, scientists take it for granted that integration must occur within the navigational system even if it is spread over different parts of the system.[16] This clearly indicates that scientists assume input-output modelling. They assume that the nervous system computes (path) integration – which mirrors the velocity-position relation of the locomotion – to keep track of the relative position of the animal.

In their review paper Etienne and Jeffery (2004) describe the information-processing function as follows:

> How is information about angular motion processed? Recently it has been found that cells in the dorsal tegmentum code for angular velocity (Sharp et al., 2001; Bassett and Taube, 2001), information they receive from the semicircular canals via the vestibular nuclei. The picture that seems to be emerging is that information about angular acceleration in the horizontal plane is collected and converted to an angular velocity signal by the semicircular canals, then passed on to the dorsal tegmentum and integrated again on its way through the mammillary nuclei and thalamus (Bassett and Taube, 2001). This provides an angular distance measure that updates the head direction signal appropriately (p. 183).

Etienne and Jeffery describe here a process with two integration steps. The inputs to the vestibular system are signals of angular acceleration; these are converted to angular velocity signals (first integral). The latter signals are then converted again into the angular distance measure (second integral). The authors describe here a double-mirroring process. In the first step, the nervous system converts input signals that encode acceleration to output signals that encode velocity by computing mathematical integration. This input-output function mirrors the acceleration-velocity relation, which is of mathematical integration. In the second step, the nervous system converts input signals that encode velocity (these are the outputs of the first step in the process) to output signals that encode position, by computing mathematical integration. This input-output function mirrors the velocity-position relation,

---

[14] This ability is achieved by different animals. A well-known example is the desert ant (*Cataglyphis fortis*) that returns home after an outward travel of hundreds of meters.

[15] See Mittelstaedt & Mittelstaedt (1982), Collett & Collett (2000), Etienne & Jeffery (2004), Conklin & Eliasmith (2005), McNaughton et al. (2006) and Gallistel and King (2009).

[16] It has been more recently suggested that path integration in rats is computed by the grid cells located in the dorsolateral medial entorhinal cortex (dMEC) (Hafting et al. 2005).

which is of mathematical integration too. Taken together, the overall input-output function of the double integral mirrors the acceleration-position relation, and this function consists of a sequence of two input-output integration functions, the first mirrors the acceleration-velocity relation whereas the second mirrors the velocity-position relation.

Locating targets in head-centered coordinates. Changing reference or coordinate frames is central to many visuo-motor tasks. Andersen et al. (1985) argue that the *posterior parietal cortex* (PPC) of macaque monkeys is home for the information-processing task of relocating a target in body-centered or head-centered coordinates. Experimental results indicate that that the PPC includes three types of cells: (1) Cells that respond to eye position only (15% of the sampled cells); (2) Cells that are not sensitive to eye orientation (21%), but have an activity field in retinotopic coordinates; (3) Cells that combine information from retinotopic coordinates with information about eye orientation (57%).

Zipser and Andersen (1988) hypothesized that the PPC combines retinotopic and extraretinal (eye-orientation) signals in order to compute target location in head-centered coordinates. They trained a neural network with the aim of simulating this computation (fig. 5). They used a three-layer network in which the two sets of input units model the behavior of the first two groups of cells, (1) and (2). The input layer projects to a layer of hidden units, which aims to model the activity of the third group of cells, (3). The output units encode the target's position in head-centered coordinates; cells with this property were not found in the PPC. Zipser and Andersen's impressive result is that the activity of the *hidden units*, after the training period, is very similar to the response properties of the third-group cells that combine information about eye orientation and the target's retinotopic location. Given this result, Zipser and Andersen hypothesized that there are head-centered target-location cells somewhere in the brain, cells that are the correspondents of the output units on the network model.

Rick Grush (2001), who analyzed this model, refers to the computations by the third-group PPC cells as follows: "We can suppose that the function computed by an idealized posterior parietal neuron is something like $f = (e\text{-}e_P)\sigma(r\text{-}r_i)$" (p. 161). He also notes, however, that this mathematical equation applies to two different relations. It refers to the neural relation between the two groups of "input" PPC cells. The activity of the "output" PPC cells (group (3)) is a multiplication of the activity of the groups (1) and (2). But the mathematical

equation also refers to some complex relations between what is being represented. What is being represented by the output, which is the "stimulus distance from preferred direction relative to the head" (p. 161) is a multiplication of the properties encoded by the inputs, namely, the difference between actual and preferred eye orientation ($e-e_P$) and (gaussian of) the distance from the retinal location of stimulation from the receptive field ($\sigma(r-r_i)$). So we see again that there is a morphism relation between the nervous system and the world. The input-output function (of multiplication) preserves a pattern of relation – between eye orientation and stimulus retinotopic location – that can also be described in terms of multiplication.

To sum up, we looked at three central works in computational neuroscience and saw that in all of them the nervous system is described as an input-output model of a target domain. In some cases, the mirroring is more apparent, whereas in other cases, it takes some effort to make it explicit. In yet other cases, it might not be explicated at all. Can we generalize from three examples to computational theories in cognitive neuroscience more generally? I cannot demonstrate that the input-output modelling assumption is held everywhere, but I think that it is very widespread and very central in computational theories of cognition. In what follows I will support this claim by showing that the morphism relation plays a key theoretical role – both methodological and explanatory – in cognitive neuroscience.

3. Methodological role

Input-output modelling plays an important methodological role in discovering the input-output function that the nervous system computes. In many cases, environmental cues are used to infer the computed function. Input-output modelling has a key role in this inference. Consider our oculomotor system. Scientists discovered that the inputs to the system are velocity signals. They also hypothesized that these signals are translated to position signals that are crucial to move the eyes to new positions. Assuming that the velocity-position relation is that of integration, they inferred that there is a sub-system that performs this transformation by computing integration. They thus called the system the *neural integrator*.

We can put the inference, somewhat crudely, as follows:

- Electrophysiological experiments show that input cells encode eye velocity. Other (output) cells encode eye position.

- The eye's velocity-position relation, in the *target domain*, is that of mathematical integration.
- Therefore: The input-output function computed by the neural system is integration.

But one can notice that the conclusion does not follow from the two premises. Why infer that the *inner* function is that of integration from the premise that the *outer* function is that of integration? The inference becomes valid *if* we also assume that the (inner) input-output function mirrors the velocity-position relation. When making the additional (third) premise the argument looks as follows:

- Electrophysiological experiments show that input cells encode eye velocity. Other (output) cells encode eye position.
- The eye's velocity-position relation, in the *target domain*, is that of mathematical integration.
- *The computed input-output function mirrors the eye's velocity-position relation.*
- Therefore: The input-output function computed by the neural system is integration.

The advantage of this methodology is that we can learn about the inner function of the nervous system, which is often hidden and hard to decipher, from the outer function that is often apparent. This is not the end of the scientific investigation, of course. Further studies are conducted in order to confirm the conclusion and to locate the integrator in the nervous system. More studies aim to characterize how the system performs integration, namely, the mechanisms that conduct the input-output transformation. The important moral, however, is that the input-output modelling assumption is entrenched in cognitive neuroscience. Theoreticians like Robinson (1989), Seung (1998) and many others are deeply convinced that there must be an "integrator" within the oculomotor system that mirrors the velocity-position relation. They take it to be obvious that if the outer relation between the represented entities is that of integration, the nervous system somewhere mirrors this relation by computing integration too.

The same goes for path integration. In the paragraph quoted above, Etienne and Jeffery take it as obvious that the computation of input signals that encode acceleration to output signals that encode velocity is that of integration. They assume, in other words, that the relevant computation mirrors the acceleration-velocity relation and, hence, must be integration. They make the same assumption about the second integral. They take it as obvious that the computation of input signals that encode velocity to output signals that encode position is that of integration. They infer, in other words, that the nervous system must compute double-integral. This inference – from outer relation to inner function – is valid under the

assumption that the nervous system is an input-output model of the animal's movement. The assumption, more precisely, is that the overall input-output function of double integral mirrors the acceleration-position relation, and that this function consists of a sequence of two input-output integration functions – the first mirrors the acceleration-velocity relation whereas the second mirrors the velocity-position relation. Without this assumption Etienne and Jeffery cannot reach their conclusion that the system computes double-integration. Again, this assumption of input-output modelling is not made explicitly. It is an implicit assumption about our brain-world relations that underlies the scientific investigation.

We see, then, that the methodology of discovering the input-output function from outer relations in the target system is fairly common in cognitive neuroscience. But it is certainly not the only way to discover the computed function. When scientists don't know or are unsure about the outer relation, they cannot infer about the inner function and they thus use different methodologies to discover the inner function. Thus in Zipser and Andersen (1988), we do not see a progression from the outer function to the inner function. The fact that the input-output function in the nervous system – (1) + (2) → (3) – is that of multiplication is discovered through the training of the (artificial) neural network that simulates the (real) neural computation. The modelling relation between the inner and outer relation is featured only later, in the analysis of the neural network. Zipser and Andersen might have assumed that the computation from the first two groups of cells to the third – (1) + (2) → (3) – mirrors some relation between the represented items. In this respect, the input-output modelling assumption is also featured in their work. Nevertheless, they did not know the exact nature of the mirrored relation and, hence, of the computation. They thus trained the network to find the input-output function instead.

4. Explanatory role

On the explanatory side, input-output modelling serves to address certain *why* questions. The question at focus is why a given mathematical function is relevant to, or appropriate for, the explanandum cognitive task. Consider our neural integrator. Its task is to produce codes of eye position ("output") from codes of eye velocity ("input"). The system accomplishes the task by computing mathematical integration. The question is: *Why* does computing integration lead to codes of eye position? To see the force of this question we can contrast integration with other mathematical functions. We can then ask: Why does the nervous

system compute integration rather than other mathematical functions – say, multiplication, exponentiation, or factorization – to produce codes of eye position? What makes integration appropriate for producing eye-position codes?

Inner mechanisms do not provide an answer to this question. They can certainly answer the question of *how* the function is being computed. Specifying the algorithm tells us how the input values are mapped to output values, and specifying the underlying neural structures tells us how the neural mechanism enables this computation. But the *why*-question is not about the inner mechanisms that give to the input-output function but about the brain-world relations. The question is about the relations between the inner (computed) function and the information-processing task that is defined, at least partly, by the target system, e.g., properties of the eyes (in controlling eye movement). If you remove the neural integrator – with the same algorithmic and neural mechanisms – to a very different environment, one in which the relations between the velocity and position are very different, then computing integration might no longer end up with codes of eye position. These considerations show that the why-question applies equally well to algorithmic and neural mechanisms. The question is why these algorithmic and neural mechanisms produce the explanandum cognitive phenomenon. After all, when changing the environment, the very same algorithmic and neural mechanisms no longer produce, hence cannot account for, the explanandum cognitive phenomenon.

Input-output modelling provides an answer to this *why* question. The neural network computes integration *because* integration preserves the velocity-position relation, namely, the (integration) relation between eye movement and eye position 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. The same goes for multiplication, exponentiation and many other functions. Integration, however, is appropriate for the task: 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.

Woodward (2003) famously proposes that causal information is explanatory by virtue of allowing answering what-if-things-had-been-different questions. Others have recently suggested that such information is explanatory even if it is not causal (Chirimuuta 2014;

Rusanen & Lappi 2016). Input-output modelling answers relevant what-if-things-had-been-different questions. We can see, for example, that if we intervene in input-output modelling, then computing integration is no longer appropriate for producing codes of eye position. We can interfere in input-output modelling, either by changing the inner input-output function or by changing the velocity-position relation. In neither case does the system any longer produce codes of eye position:

- If the system had not computed integration (but rather exponentiation), the system would not have produced codes of eye position.
- If the world had changed so that the eye's velocity-position relation were not integration (but exponentiation), the system (when computing integration) would not have produced codes of eye position.

Let me be a bit more precise about the nature of the explanandum and the structure of modelling explanations. Another way to present the explanandum *why*-question is as follows: The computation starts with an input neural value $\dot{E}$ that encodes some distal feature $\underline{\dot{E}}$, i.e., eye velocity. It computes a certain function f, i.e., mathematical integration, whose output is another neural value, E, that encodes another distal feature $\underline{E}$, e.g., eye position. The explanandum question is why this computation (e.g., integration), which starts from neural values that encode eye velocity, terminates in neural values that encode eye position.

To put it succinctly, the given premises are:

(P1) $I(\dot{E}) = \underline{\dot{E}}$ (the neural activity $\dot{E}$ encodes $\underline{\dot{E}}$).

(P2) $f(\dot{E}) = E$ (f maps input neural values $\dot{E}$ to output neural values E).

And the conclusion is:

(C) $I(E) = \underline{E}$.

The question, when put this way, is about the inference from (P1) and (P2) to (C). The answer is in no way trivial. If we change the environment, the same computation, f, which starts from the same velocity-coded neural input values, $\dot{E}$, will still terminate with the same neural output values E, but E might no longer encode eye position or anything at all. Why, then, does computing f (i.e., integration) over neural input values, E, that encode eye velocity end up with neural codes of eye position?

Input-output modelling asserts that, at least in our world, the input-output function mirrors the velocity-position relation. We can formulate this assertion in two assumptions. One is that the velocity-position relation, in the abstract, is that of mathematical integration too:

(P3) $f(\underline{\dot{E}}) = \underline{E}$ (the integration on velocity values, with respect to time, yields position values).

(P4) $f(I(\dot{E})) = I(E)$.

(P3) asserts that there is an input-output morphism between the modelling nervous system and the target eyes. This is in parallel to the second condition in the characterization of input-output modelling (section 1). The condition was that $f(x)=y$ iff $<\underline{x},\underline{y}>\in\underline{R}$, where f is the input-output function and $\underline{R}$ is the mirrored relation. This is tantamount to the claim that the two relations are structurally, e.g., mathematically, similar, namely, that both relations are characterized, in the abstract, by the mathematical function f (i.e., integration), which is stated by (P2) and (P3). (P4) is parallel to the third condition in the characterization of input-output modelling. The third condition states that the morphism relation coincides with the representation (or coding) relation, I, which means that $f(I(\dot{E})) = I(E)$.

From premises (P1)-(P4), we can reach the conclusion (C). Given that $I(\dot{E})= \underline{\dot{E}}$ (from (P1)) and given that $f(\underline{\dot{E}})=\underline{E}$ (from (P3)), we can infer, from (P4), that $\underline{E} = I(E)$, which is in fact the conclusion, (C).

In sum, I have argued that input-output modelling plays a role in explaining why a computed function is appropriate for the explanandum cognitive task. The appropriateness question naturally arises when we describe the nervous system as an information-processing system. The input-output modelling provides a simple answer to this question in relating the computed function to the target domain. In the last two sections I compare the modelling explanations with mechanistic (section 5) and optimality (section 6) explanations.

5. Mechanistic explanations

The mechanistic approach to explanation has been widely advocated in recent years, especially in the biological and cognitive sciences (Bechtel and Richardson, 1993; Machamer, Darden and Craver, 2000; Glennan, 2002). According to one characterization, “Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions” (Machamer et al. 2000: 3), and

mechanistic explanations describe the aspects of the mechanisms that are relevant to producing the (explanandum) phenomenon. Furthermore, it has been argued that computational models and explanations in cognitive neuroscience are mechanistic too. David Kaplan, for example, argues that "computational models in this domain [computational neuroscience] possess explanatory force to the extent that they describe the mechanisms responsible for producing a given phenomenon—paralleling how other mechanistic models explain" (2011: 339).[17]

I would like to argue that the modelling explanations outlined above are not mechanistic explanations. Whether this claim conflicts with the mechanistic framework largely depends on how strongly we interpret the claims about mechanistic explanations. Proponents of the mechanistic framework recently emphasized that they do *not* insist that all scientific explanations must be mechanistic (Craver 2016; Kaplan 2017).[18] And they can certainly concede that some explanations of cognitive phenomena have non-mechanistic, e.g., modelling, components. Given this understanding, there is no conflict between modelling and mechanistic explanations.[19]

Why are modelling explanations not mechanistic? For one thing, they are not decompositional: Explaining the appropriateness of the computed function to the cognitive task does not involve decomposition. The explanation does not decompose the input-output function into its constituents. It rather refers to the relations between the inner mechanisms and environmental features. Moreover, these relations and features are not necessarily causal. The crucial premise is a morphism relation between the inner input-output function and an outer relation in the target system, whereas the outer relation need not be causal (e.g., velocity-position relation). Another premise in the explanation is the encoding or representation relation that might not be fully explicated in mechanistic terms.

---

[17] See also Kaplan and Craver (2011), Piccinini and Craver (2011), Miłkowski (2013) and Boone and Piccinini (2016).

[18] They also point out that a "full-blown" mechanistic explanation need not specify the entire properties of the mechanism. It should specify the entire properties that are *relevant* to the explanandum phenomenon; in some cases (e.g., computational explanations) these properties might all be abstract (e.g., medium-independent) properties (Boone and Piccinini 2016).

[19] There is tension, however, about what counts as a computational explanation. Kaplan seems to claim that computational explanations in neuroscience are adequate to the extent that they describe relevant mechanisms (see also Piccinini 2015; Miłkowski 2013). We suggest that computational explanations of information-processing phenomena also involve a modelling, non-mechanistic, component (Bechtel and Shagrir 2015; Shagrir and Bechtel 2017).

Someone might say that the modelling explanation is a *sketch* of mechanism. A sketch is a description of a mechanism in which some relevant structural properties are missing. Once these missing properties are filled in, the description turns into "a full-blown mechanistic explanation"; the sketches themselves can be thus seen as "elliptical or incomplete mechanistic explanations" (Piccinini and Craver 2011: 284). They are a guide or a first step towards the full-blown mechanistic explanations. In replying, I want to emphasize that I do not deny that the modelling relations are implemented by mechanisms; this, however, does not make modelling explanations sketches, as long as the missing implementational details are not relevant to the explanandum phenomena (Shapiro 2016). I also do not deny that the mechanisms that underlie the computed function – the algorithmic and perhaps even the neuronal – are part of the complete explanation of the cognitive phenomenon. These implementational details, however, are not essential to addressing the *why* question, of why the system computes this function; they are rather relevant to addressing the question of how the function is being computed.

Moreover, specifying the mind-world implementational relation does not seem to be essential to the explanandum cognitive phenomenon. At least we see very little effort, if any, to describe these mechanisms in computational theories of cognition. Take our oculomotor integrator. We can assume that the mirroring relation between the integration function and the outer, velocity-position relation, was established in a very long evolutionary process whose result is the mirroring integrator (we can also suppose that different evolutionary mechanisms take place in different species such as goldfish, cats and primates). But no one appeals to these implementational processes in order to explain the appropriateness of inner functions and mechanisms to the cognitive task. This, in my view, is not too surprising: The relevant feature for the explanation is the morphism relation between the mirroring integrator and the velocity-position relation and not the details about the implementation of the mapping. When we see that these modelling relations are in place we can understand why the inner neural mechanisms (including the computed input-output function) end up in codes of eye position. The mind-world evolutionary story might explain how those modelling relations were established in the first place, but they add little, if anything, to the conclusion that computing integration is appropriate to producing codes of eye position.

I do not say that the modelling explanation is the "only game in town", and that it might not be replaced in future by a full-blown mechanistic explanation. My point is that the modelling assumption is currently entrenched in theoretical and computational approaches in cognitive neuroscience, and for a reason. The question of appropriateness arises when we describe the brain as an information-processing system. When we explain an information-processing task that is partly defined in terms of what is being represented, we are forced to address the question of why the processes that take place inside the brain, which start from some input representations, lead to the appropriate output representations. Input-output modelling provides an explanation for this appropriateness. Whether modelling explanations will be replaced in the future by full-blown mechanistic explanations is yet to be seen, but is certainly not to be ruled out.

6. Optimality

In a recent paper, Mazvita Chirimuuta (2014) introduces the notion of *I*-minimal models. These computational models are minimal in the sense that "they typically abstract away from many biophysical details of the neural system" (p. 128).[20] My focus here is on the *I*-aspect of *I*-minimal, which alludes to *interpretative models* (Dayan and Abbott 2001). Dayan and Abbott note that theoretical neuroscience invokes, in addition to phenomenal (descriptive) and mechanistic models, interpretational models. These models "use computational and information-theoretic principles to explore the behavioral and cognitive significance of various aspects of nervous system function, addressing the question of why nervous systems operate as they do" (2001:1).

At first glance, it seems that Chirimuuta (via Dayan and Abbott) and I are alluding to the same *why* questions, of why nervous systems operate as they do. But the answers go in different directions. I argue that answering these *why* questions involves input-output modelling. Chirimuuta argues that answering *why* questions about the operations of nervous

[20] Chirimuuta argues that this minimality conflicts with the more chauvinistic statements about the dominance of mechanistic explanations. Talking about the normalization model, she says that "my key claim is that the use of the term 'normalization' in neuroscience retains much of its original mathematical-engineering sense. It indicates a mathematical operation—a computation—not a biological mechanism", and that this model "departs fully from the model-to-mechanism mapping framework that has been proposed as the criterion for explanatory success" (Chirimuuta 2014); she refers here to Kaplan's model-to-mechanism mapping (3M) requirement (Kaplan 2011; Kaplan and Craver 2011). For a reply see Kaplan (2017) who argues that the implementation of the normalization equation (in different species) is an essential part of the explanation.

systems invokes explanations ("interpretative models") that typically make reference to efficient coding principles. Her main example is the normalization equation that models the cross-orientation suppression of simple cell response in the primary visual cortex and in other systems. Very briefly, while cells in V1 were found to selectively respond to bar-shaped stimuli in a preferred orientation (Hubel and Wiesel 1962), it turns out that this response is significantly reduced ("suppressed") if the preferred stimuli are super-imposed by other stimuli with different, non-preferred, orientation. Heeger (1992) advanced the normalization model to account for the phenomenon. The idea is that in addition to the excitatory input from LGN, each V1 cell also receives inhibitory inputs from its neighboring V1 cells (that are sensitive to lines in different angles). As Chirimuuta emphasizes, this normalization equation – which quantitatively describes the cells' responses – is later found in other parts of the nervous system (Carandini and Heeger 2012). This raises the question: "*why should so many systems exhibit behavior described by normalization equation?*" And the answer to this is that "*for many instances of neural processing individual neurons are able to transmit more information if their firing rate is suppressed by the population average firing rate*" (p. 143).

How does this account of *why* questions, in terms of efficient coding principles, comport with my account of *why* questions, in terms of modelling? My tentative answer is that the accounts are different as the *why* questions are different.[21] But the questions are not disconnected. My concern is questions of the form: Why is a certain function *f* appropriate (or not) for a certain task? Chirimuuta is concerned with the further question: Take all the functions f_1, f_2, … that are appropriate for the task. Why choose f_i rather than the other *f*'s?[22]

To see the difference, take the example of edge-detection. Marr (1982; Marr and Hildreth 1980) argues that V1 cells detect edges by computing the zero-crossings of second-derivative Laplacian operators. The latter operators are applied by the ganglions and LGNs to the retinal image and are described, quantitatively, by the formula $\nabla^2 G * I$, where *I* is the image, $*$ is a convolution operator and $\nabla^2 G$ is a filtering operator: G is a Gaussian that blurs the image, and ∇^2 is the Laplacian ($\partial^2/\partial x^2 + \partial^2/\partial y^2$). One question we can ask, as Marr does, is why this computation is appropriate for detecting edges.[23] The answer, I suggested, is

[21] Colin Klein suggested that we might be dealing here with different *why* questions.
[22] A similar question arises for the different algorithms that support the same function, which is *why* using one algorithm rather than another.
[23] Marr writes:

provided in terms of modelling (Shagrir 2010; Shagrir and Bechtel 2017). In particular, this input-output function preserves sharp changes in reflectance and illumination in the visual field that happen to occur along physical edges (e.g., object boundaries), and that can be described in terms of derivation. Other functions – factorization, exponentiation, division – that do not preserve the pertinent relation are obviously not appropriate for edge-detection.

But, now, there are other functions that might be appropriate for the task too. As Marr noticed, the visual system could detect edges by computing the extreme points of first-derivative operators, the second-order directional derivatives and perhaps other appropriate functions. So there is a further question: Why compute the zero-crossing of *second-derivative Laplacian operators* rather than computing other derivative (directional) operators, which are appropriate too. I think that Chirimuuta is concerned with this further question. Assuming that the task is responding to oriented lines ("edges"), her question is: Why compute the normalization equation (cross orientation suppression) rather than (say) a simple linear response to the receptive-field properties. In many cases, the answer has to do with the efficiency of computation. Given that there is a limit to the amount of information-processing possible in the brain, the expected simple-linear-response function might not be consistent with the limitations of the brain. In this case, we appeal to efficient-coding principles and other canons of information theory. Indeed, Marr discusses this point of efficiency in some detail in his theory of edge-detection (1982: p. 56 ff.). He writes that "the great advantage of using it [Laplacian operator] is economy of computation" (p. 56). The computation of the directional derivative operators is costly, whereas using the Laplacian operators is efficient and satisfactory.

My tentative proposal, then, is that computational theories of cognition are concerned with a family of *why* questions about the operations of the nervous system. Some questions are about the appropriateness of these operations to the cognitive tasks. Other questions address the advantage of these operations over other (seemingly) appropriate operations, and there might be other kinds of *why* questions as well.

> Up to now I have studiously avoided using the word *edge*, preferring instead to discuss the detection of intensity changes and their representation by using oriented zero-crossing segments. The reason is that the term *edge* has a partly physical meaning – it makes us think of a real physical boundary, for example – and all we have discussed so far are the zero values of a set of roughly band-pass second-derivative filters. We have no right to call these edges, or, if we do have a right, then we must say so and why. (1982: 68)

7. Summary

I argued that input-output modelling is central to computational work in cognitive neuroscience. Looking at three examples, we saw that in some cases the modelling is more apparent (as in the examples of integration), whereas in other cases it takes more effort to expose the modelling relation (as in changing reference framework). I then explicated the central theoretical role of input-output modelling. It plays a methodological role, in discovering what function is being computed. And it plays an explanatory role, in accounting for the appropriateness of the computed function for the explanandum cognitive task. Finally, I compared very briefly the modelling explanation to mechanistic and optimality explanations, noting that in both cases the explanations can be seen as complementary rather than contrastive or competing. I haven't discussed the role of input-output modelling in the characterization of representation and computation. I leave this endeavor for another occasion.

Acknowledgements: I am grateful to Lotem Elber-Dorozko, Jens Harbecke, Shahar Hechtlinger, David Kaplan, Colin Klein, Arnon Levy, Gal Patel and two anonymous referees for their comments. Early versions of the paper were presented at seminars in Macquarie University, Tel-Aviv University, University of Canterbury, University of Otago and at the following conferences: *The Aims of Brain Research: Scientific and Philosophical Perspectives* (Jerusalem), *Conference of the International Association for Computing and Philosophy* (Thessaloniki), and *the 7th AISB Symposium on Computing and Philosophy* (London). I thank the participants for stimulating discussion. This research was supported by a grant from GIF, the German-Israeli Foundation for Scientific Research and Development.

References

Andersen, R.A., Essick, G.K., & Siegel, R.M (1985). Encoding of spatial location by posterior parietal neurons. *Science* 230: 456–458.

Bartels, A. (2006). Defending the structural concept of representation. *THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia* 21: 7–19.

Bassett J.P., & Taube, J.S. (2001). Neural correlates for angular head velocity in the rat dorsal tegmental nucleus. *Journal of Neuroscience* 21: 5740–5751.

Bechtel, W., (2012). Understanding endogenously active mechanisms: A scientific and philosophical challenge. *European Journal for Philosophy of Science* 2: 233–248.

Bechtel, W., & Richardson, R. C. (1993). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton University Press.

Bechtel, W., & Shagrir, O. (2015). The non-redundant contributions of Marr's three levels of analysis for explaining information-processing mechanisms. *Topics in Cognitive Science* 7: 312–322.

Boone, W., & Piccinini, G. (2016). Mechanistic abstraction. *Philosophy of Science* 83: 686–697.

Cannon, S.C. and Robinson, D. (1987). Loss of the neural integrator of the oculomotor system from brain stem lesions in monkey. *Journal of neurophysiology* 57: 1383-1409.

Carandini, M., & 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.

Churchland, P. M. (2007). *Neurophilosophy at Work*. Cambridge University Press.

Clark, A. (2015). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Collett, M., & Collett, T.S. (2000). How do insects use path integration for their navigation? *Biological Cybernetics* 83: 245–259.

Conklin, J., & Eliasmith, C. (2005). Controlled attractor network model of path integration in the rat. *Journal of Computational Neuroscience* 18: 183–203.

Craver, C.F. (2016). The explanatory power of network models. *Philosophy of Science* 83: 698–709.

Cummins, R. (1989). *Meaning and Mental Representation*. MIT Press.

Da Costa, N.C.A., & French, S. (2003). *Science and Partial Truth: A Unitary Understanding of Models and Scientific Reasoning*. Oxford University Press.

Dayan, P., & Abbott, L.F. (2001). *Theoretical Neuroscience: Computational and Mathematical*

Modeling of Neural Systems. MIT Press.

Eliasmith, C., & Anderson, C.H. (2003). *Neural Engineering: Computation, Representation and Dynamics in Neurobiological Systems*. MIT Press.

Etienne, A.S., & Jeffery, K.J. (2004). Path integration in mammals. *Hippocampus* 14: 180–192.

Fodor, J.A. (1994). *The Elm and the Expert: Mentalese and Its Semantics.* MIT Press.

French, S., & Ladyman, J. (1999). Reinflating the semantic approach. *International Studies in the Philosophy of Science* 13: 103-121.

Frigg, R., & Hartmann, S. (2017). Models in science. In Zalta E.N. (ed.), *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/entries/models-science>.

Gallistel, C.R., & King, A. (2009). *Memory and the Computational Brain: Why Cognitive Science will Transform Neuroscience*. Blackwell/Wiley

Giere, R.N. (2004). How models are used to represent reality. *Philosophy of Science* 71: 742–752.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science* 69: S342–S353.

Glimcher, P.W. (1999). Oculomotor control. In R.A. Wilson and F.C. Kiel (eds.) *MIT Encyclopedia of Cognitive Science* (pp. 618–620). MIT Press.

Goldman, M.S., Kaneko, C.R., Major, G., Aksay, E., Tank, D.W., & Seung, H.S. (2002). Linear regression of eye velocity on eye position and head velocity suggests a common oculomotor neural integrator. *Journal of Neurophysiology* 88: 659-665.

Griffiths, T.L., Kemp, C., & Tenenbaum, J.B. (2008). Bayesian models of cognition. In R. Sun (ed.), *The Cambridge Handbook of Computational Cognitive Modleing* (pp. 59–100). Cambridge University Press.

Grush, R. (2001). The semantic challenge to computational neuroscience. In Machamer, P., Grush, R., & McLaughlin, P. (eds.), *Theory and Method in the Neurosciences* (pp. 155–172). University of Pittsburgh Press.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences* 27: 377–442.

Hafting, T., Fyhn, M., Molden, S., Moser, M-B., & Moser, E.I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature* 436: 801–806.

Haugeland, J. (1981). Semantic engines: An introduction to *Mind Design*. In Haugeland, J. (ed.), *Mind Design: Philosophy, Psychology, and Artificial Intelligence* (pp. 1–34). MIT Press.

Heeger, D.J. (1992). Normalization of cell responses in cat striate cortex. *Visual neuroscience* 9: 181-197.

Hubel, D.H., & Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology* 160: 106-154.

Kaplan, D.M. (2011). Explanation and description in computational neuroscience. *Synthese*

183: 339–373.

Kaplan, D.M. (2017). Neural computation, multiple realizability, and the prospects for mechanistic explanation. In Kaplan, D.M. (ed.), *Explanation and Integration in Mind and Brain Science*. Oxford University Press (forthcoming).

Kaplan, D.M., & Craver, C.F. (2011). The explanatory force of dynamical and mathematical models in neuroscience : A mechanistic perspective. *Philosophy of Science* 78: 601–627.

Leigh, R.J., & Zee, D.S. (2006). *The Neurology of Eye Movements* (4th ed.). Oxford University Press.

Machamer, P., Darden, L., & Craver, C.F. (2000). Thinking about mechanisms. *Philosophy of Science* 67: 1–25.

McNaughton, B.L., Battaglia, F.P., Jensen, O., Moser. E.I., & Moser, M-B. (2006). Path integration and the neural basis of the 'cognitive map'. *Nature Reviews Neuroscience* 7: 663-678.

Marr, D.C. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman.

Marr, D.C., & Hildreth, E.C. (1980). Theory of edge detection. *Proceedings of the Royal Society of London, Series B, Biological Sciences* 207: 187–217.

Miłkowski, M. (2013). *Explaining the Computational Mind*. MIT Press.

Mittelstaedt, H., & Mittelstaedt, M-L. (1982). Homing by path integration. In Papi, F., & Wallraff, H.G. (eds.), *Avian navigation* (pp. 290–297). Springer.

O'Brien, G., & Opie, J. (2009). The role of representation in computation. *Cognitive Processing* 10: 53–62.

O'Keefe, J., & Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Clarendon Press.

Piccinini, G. (2015). *Physical Computation: A Mechanistic Account*. Oxford University Press.

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

Pylyshyn, Z.W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. MIT Press.

Ramsey, W. (2007). *Representation Reconsidered*. Cambridge University Press.

Robinson, D.A. (1968). The oculomotor control system: A review. *Proceedings of the IEEE* 56: 1032-1049.

Robinson, D.A. (1989). Integrating with neurons. *Annual Review of Neuroscience* 12: 33–45.

Rusanen, A-M., & Lappi, O. (2016). On computational explanations. *Synthese* 193: 3931–3949.

Ryder, D. (2004). *SINBAD* neurosemantics: A theory of mental representation. *Mind & Language* 19: 211–240.

Seung, H.S. (1998). Continuous attractors and oculomotor control. *Neural Networks* 11: 1253–1258.

Shagrir, O. (2010). Marr on computational-level theories. *Philosophy of Science*: 477–500.

Shagrir, O. (2012), Structural representations and the brain. *British Journal for the Philosophy of Science* 63: 519–545.

Shagrir, O., & Bechtel, W. (2017). Marr's computational level and delineating phenomena. In Kaplan, D.M. (ed.), *Integrating Mind and Brain Science: Mechanistic Perspectives and Beyond*. Oxford University Press (forthcoming).

Shapiro, L.A. (2016). Mechanism or bust? Explanation in psychology. *British Journal for the Philosophy of Science* (forthcoming).

Sharp, P.E., Tinkelman A., & Cho, J. (2001). Angular velocity and head direction signals recorded from the dorsal tegmental nucleus of Gudden in the rat: Implications for path integration in the head direction cell circuit. *Behavioral Neuroscience*. 115: 571–588.

Suárez, M. (2010). Scientific representation. *Philosophy Compass* 5: 91–101.

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese* 87: 449–508.

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation.* Oxford University Press.

Zipser, D., & Andersen, R.A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, 331: 679–684.

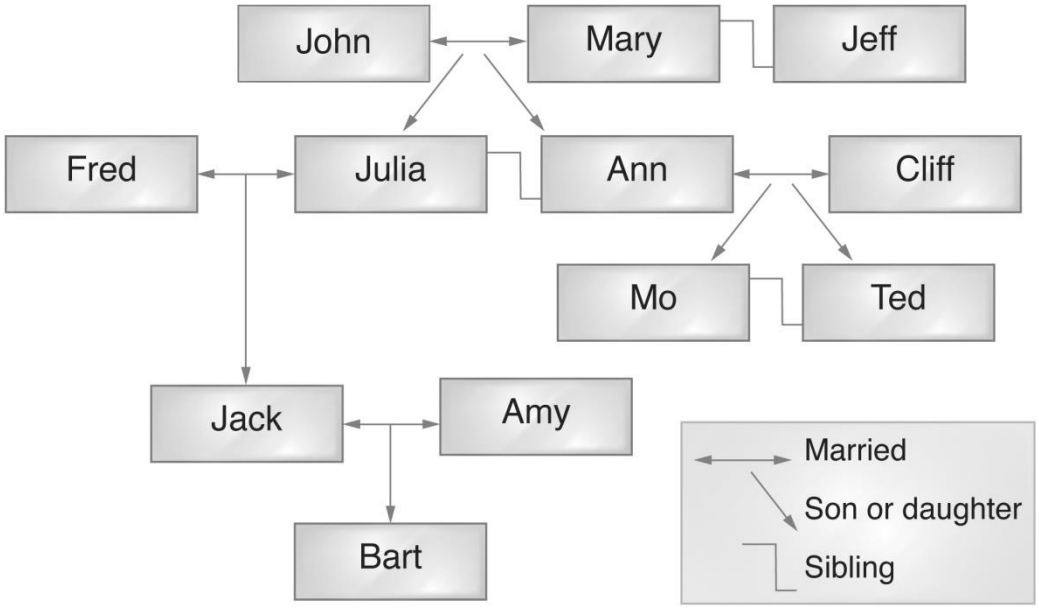

Figure 1: A family-tree model for determining familial links (from Ramsey 2007: 81 (fig. 3c); with permission from Cambridge University Press).

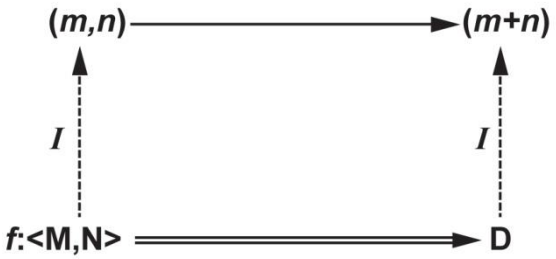

Figure 2: Cummins's London-Tower-Bridge. The bottom span is the input-output function satisfied (or computed), *f*; the double-dashed arrow is a causal process by which the system satisfies (computes) *f* which operates on numerals. The top span is the function *plus* which is defined over numbers. The mapping ("interpretation") function, *I*, maps the inputs and outputs of *f* to the inputs and outputs of *plus*. The input-output function *f* thus mirrors or preserves the *plus* function.

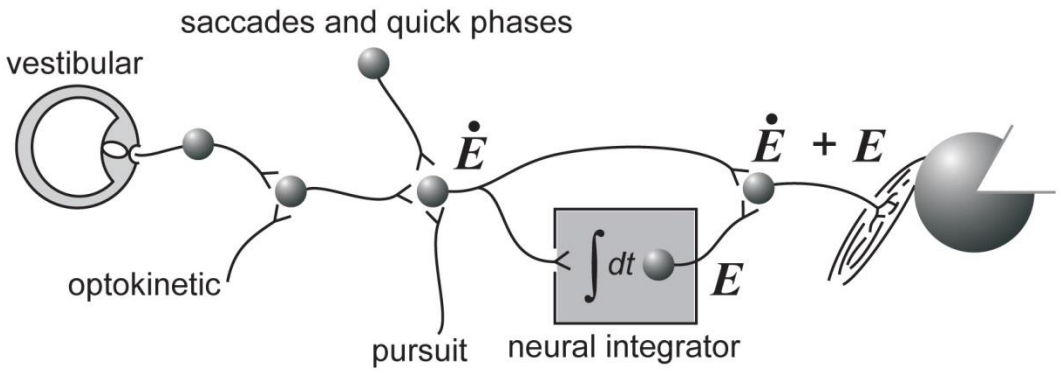

Figure 3: The common neural integrator. The neural integrator receives as inputs eye-velocity encoded signals, *Ė*, and produces eye-position encoded outputs *E*. The velocity codes, *Ė*, combine the vestibular, optokinetic, saccadic, and pursuit velocities. These codes are projected directly to the motoneurons that produce velocity commands. These codes are also projected to the neural integrator that produces position codes, which are in turn projected to the motoneurons for position commands (Adapted from Cannon and Robinson 1987: 1384 (fig. 1); reprinted by permission of the American Physiological Society (APS)).

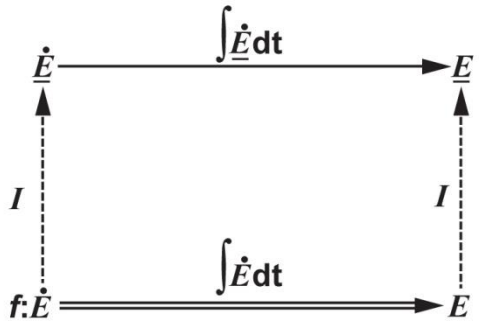

Figure 4: The oculomotor integrator as an input-output model. The lower span describes a causal process in the neural system (i.e., in the neural integrator) that transforms input values, $\dot{E}$ to output values E. The computed function, f, is mathematical integration (in the abstract), namely, the values E, are the result of mathematical integration over $\dot{E}$ with respect to time. The upper span describes the target domain, namely, the eyes. The term $\underline{\dot{E}}$ describes the velocity of the eye, whereas the term $\underline{E}$ describes the (horizontal) distance from previous eye position, namely, new eye position. The velocity-position relation is also that of integration. Thus the mapping, I, is a morphism relation, and the integrator (assuming that the inputs and outputs encode velocities and positions respectively) is an input-output model.

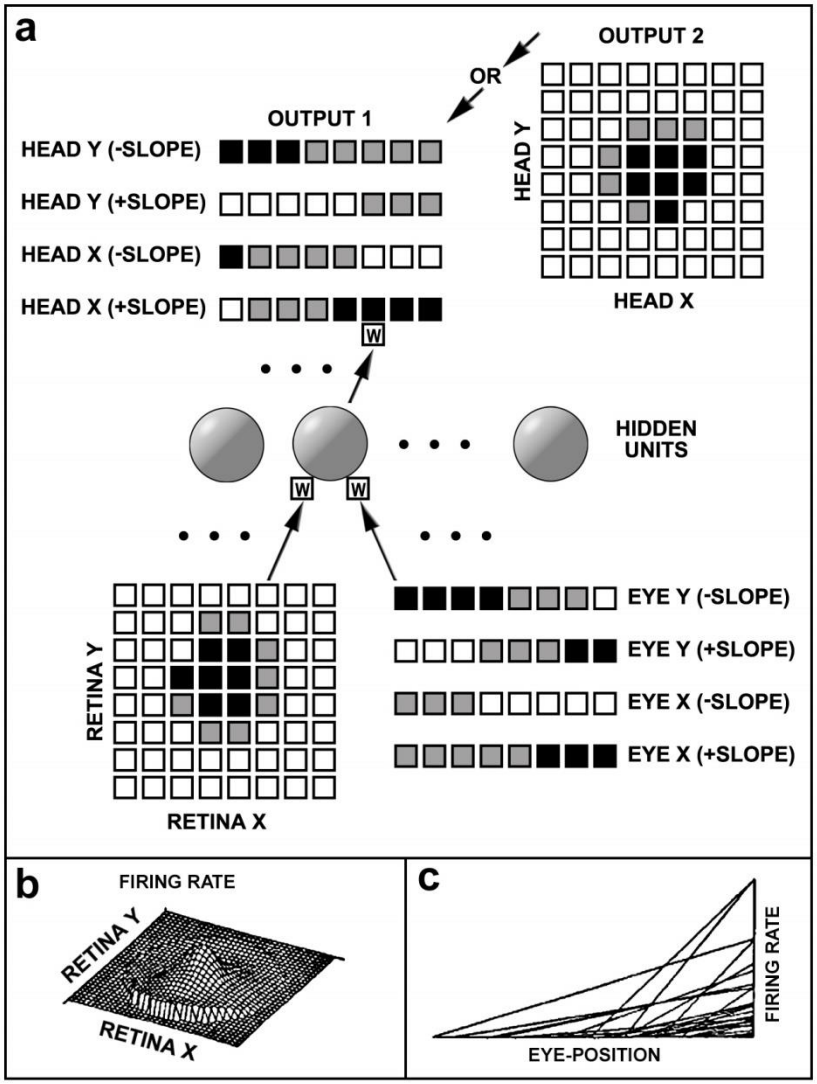

Figure 5: The Zipser-Andersen model. (a) The three-layer network, where the two sets of input units stand for the retinotopic location cells (bottom left) and eye-orientation cells (bottom right). The hidden units are meant to model the behavior of the third-group of PPC cells. The units of the output layer (two versions) stand for cells that encode the head-centered location. The network is trained through a supervised learning technique. (b) Area 7a visual neuron receptive field with a single peak near the fovea. (c) A composite of 30 area 7a-eye-position units, whose firing rates are plotted as a function of horizontal or vertical eye deviation (Reprinted by permission of Macmillan Publishers Ltd: D. Zipser & R.A. Andersen, "A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons," *Nature*, 331: 679–684 (fig. 4), copyright (1988).