Function-Theoretic Explanation and Neural Mechanisms

Frances Egan

A common kind of explanation in cognitive neuroscience might be called *function-theoretic*: with some target cognitive capacity in view, the theorist hypothesizes that the system computes a well-defined function (in the mathematical sense) and explains how computing this function contributes to the exercise of the cognitive capacity. Recently, proponents of the so-called 'new mechanist' approach in philosophy of science have argued that a model of a cognitive capacity is explanatory only to the extent that it reveals the causal structure of the mechanism underlying the capacity. If they are right, then a cognitive model that resists a transparent mapping to known neural mechanisms fails to be explanatory. I argue that a function-theoretic characterization of a cognitive capacity *can* be genuinely explanatory even absent an account of how the capacity is realized in neural hardware.

1. *Function-theoretic explanation*

Marr's (1982) theory of early vision explains edge detection by positing the computation of the Laplacian of a Gaussian of the retinal array. The mechanism takes as input intensity values at points in the image and calculates the rate of intensity change over the image. In other words, it computes a particular smoothing function. Ullman (1979) hypothesizes that the visual system recovers the 3D structure of moving objects by computing a function from 3 distinct views of 4 non-coplanar points to the unique rigid configuration consistent with the points. Shadmehr and Wise's (2005) computational account of motor control explains how a subject is able to grasp an object in view by computing the dis-

placement of the hand from its current location to the target location, i.e. by computing vector subtraction. In a well-known example from animal cognition, Gallistel (1990) explains the Tunisian desert ant's impressive navigational abilities by appeal to the computation of the displacement vector to its nest from any point along its foraging trajectory. And Sueng et al (1996, 2000) explain how the brain keeps track of eye movements by positing an internal integrator.

These examples illustrate an explanatory strategy that is pervasive in computational cognitive science. I call the strategy *function-theoretic explanation* and the mathematical characterization that is central to it *function-theoretic characterization*.[1] (Henceforth, I shall abbreviate 'function-theoretic' as FT.) Theories employing the strategy explain a cognitive capacity by appeal to an independently well-understood mathematical function under which the physical system is subsumed. In other words, what gets computed, according to these computational models, is the value of a mathematical function (e.g. addition, vector subtraction, the Laplacian of a Gaussian, a fast Fourier transform) for certain arguments for which the function is defined. For present purposes we can take functions to be mappings from sets (the arguments of the function) to sets (its values). Ideally, the theory will go on to propose an algorithm by which the computation of the value of the function(s) is effected, and describe the neural hardware that implements the computation.[2]

---

[1] This sense of *function-theoretic* characterization is not to be confused with various notions of functional explanation in the literature, in particular, with Cummins' (1975) notion of *functional analysis*. However, a functional analysis of a complex system may involve function-theoretic characterization, in the sense explicated in this paper.

[2] The FT characterization, the specification of the algorithm, and the neural implementation correspond, very roughly, to Marr's three levels of description – the computational, algorithmic, and implementation – respectively.

A function-theoretic description provides a domain-general, environment-neutral characterization of a mechanism. It prescinds not only from the cognitive capacity that is the explanatory target of the theory (vision, motor control, etc.), but also from the environment in which the capacity is normally exercised. In fact, the abstract nature of the FT characterization – in particular, the fact that as an independently characterized mathematical object the function can be decoupled from both the environmental context and the cognitive domain that it subserves – accounts for perhaps the most significant explanatory virtue of function-theoretic characterization. The mathematical functions deployed in computational models are typically well understood independently of their use in such models. Laplacian of Gaussian filters, fast fourier transforms, vector subtraction, and so on, are standard items in the applied mathematician's toolbox. To apply one of these tools to a biological system – to subsume the system under the mathematical description – provides a measure of understanding of what might otherwise be a heterogeneous collection of input-output pairs. (Aha! It's an integrator!)[3] And since the FT characterization specifies the function intensionally, typically in terms of an algorithm for computing the function, it provides the basis for predicting the output of the device in a wide range of circumstances that go well beyond the observed data set.

But, of course, the theorist of cognition must explain how computing the value of the specified function, in the subject's normal environment, contributes to the exercise of the cognitive capacity that is the explanatory target of the theory – for the motor control mechanism, the capacity to grasp an object in nearby space, for visual mechanisms, the

[3] Moreover, theorists typically have at their fingertips various algorithms for computing these functions. Of course, the algorithms are hypotheses that require independent support, but the point is that the theorist has a ready supply of such hypotheses.

capacity to see 'what is where' (as Marr puts it) in the nearby environment. Only in *some* environments would computing the Laplacian of a Gaussian help an organism to see. In our environment this computation produces a smoothed output that facilitates the detection of sharp intensity gradients across the retina, which, when these intensity gradients co-occur at different scales, correspond to physically significant boundaries – changes in depth, surface orientation, illumination, or reflectance – in the scene. Ullman's structure-from-motion mechanism succeeds in recovering the 3D structure of a moving object by computing the unique rigid configuration consistent with 3 distinct views of 4 non-coplanar points on the object only because, in *our* world, most objects are rigid in translation (the *rigidity assumption*). Thus, to yield an explanation of the target cognitive capacity, the environment-neutral, domain-general characterization given by the FT description must be supplemented by environment-specific facts that explain how computing the value of the specified mathematical function, in the subject's normal environment, contributes to the exercise of the target cognitive capacity.

One way to connect the abstract FT characterization to the target cognitive capacity is to attribute representational contents that are appropriate to the relevant cognitive domain. Theorists of vision construe the mechanisms they posit as representing properties of the light, e.g. light intensity values, changes in light intensity, and, further downstream, changes in depth and surface orientation. The inputs and outputs of the Laplacian/Gaussian filter represent light intensities and discontinuities of light intensity respectively. Theorists of motor control construe the mechanisms they posit as representing positions of objects in nearby space and changes in joint angles. But the fact that a mechanism characterized function-theoretically can *also* be characterized in terms of rep-

resentational contents appropriate to the cognitive domain in question does not obviate the explanatory interest of the more abstract, domain-general, mathematical characterization that is the focus of this paper.[4]

I will have much more to say about function-theoretic explanation as we progress, but I have said enough to set up the main issue of the paper. I turn now to the challenge from the new mechanists.

2. *The new mechanistic philosophy*

A *mechanism* is an object that performs some function in virtue of the operations of its component parts and their organization. *Mechanistic explanation* is the explanation of the capacities of a system by reference to the properties and operations of its component parts and their causal organization.[5]

Proponents of the new mechanistic philosophy claim that all genuine explanation – at least all genuine explanation in cognitive neuroscience – is mechanistic explanation:

> The common crux of mechanistic explanation, both in its current form and in forms stretching back through Descartes to Aristotle, is to reveal the causal structure of a system. Explanatory models are counted as good or bad to the extent that they capture, even dimly at times, aspects of that causal structure. (Piccinini & Craver 2011, 292)

---

[4] In Egan (forthcoming) I argue that representational contents are best construed as part of an explanatory *gloss* on a computational theory, that they serve a variety of pragmatic purposes but are, strictly speaking, theoretically optional. The argument in this paper does not depend on any particular view of representational content.

[5] For characterization of mechanisms and mechanistic explanation see Bechtel 2006, Bechtel and Abrahamsen 2005, Bechtel and Richardson 1993, Craver 2006, 2007, and Machamer, Darden, and Craver 2000.

> … explanations in computational neuroscience are subject to precisely these same norms [the norms of mechanistic explanation]. The cost imposed by departing from this view… is the loss of a clear distinction between computational models that genuinely explain how a given phenomenon is actually produced versus those that merely describe how it might possibly be produced... And it is precisely by adhering to this distinction (along with a distinction between merely describing or saving a phenomenon and explaining it), that one can identify models in computational neuroscience possessing explanatory force. (Kaplan 2011, 346)

Levy (2013) provides a nice gloss on the view that he calls *explanatory mechanism*: "…to understand a phenomenon one must look under the hood and discern its underlying structure" (p.11 pdf).[6]

The idea that a model has explanatory force just to the extent that it reveals the causal structure of an underlying mechanism is explicated in terms of what Kaplan calls a *model-mechanism-mapping (3M) constraint* on explanatory models:

> (3M) A model of a target phenomenon explains that phenomenon to the extent that (a) the variables in the model correspond to identifiable components, activities, and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the (perhaps mathematical) dependencies posited among these (perhaps mathematical) variables in the model correspond to causal relations among the components of the target mechanism. (Kaplan 2011, 347; see also Kaplan & Craver 2011, 611)

The 3M Constraint is claimed to distinguish genuine explanations in cognitive neuroscience from mere descriptions and predictive devices (Kaplan 2011, 340).

---

[6] Neither Levy nor Bechtel endorse the claim that all genuine explanation is mechanistic.

There is no doubt that many explanatory models in cognitive neuroscience do conform to the new mechanists' strictures. But if the 3M Constraint is indeed a necessary condition on genuine explanation, then many promising cognitive models turn out not to be explanatory. In the next two sections I argue that FT models often do not fit the mechanists' account, yet they can be, and often are, genuinely explanatory.

3. *Function-theoretic models and mechanistic explanation*

Satisfying the 3M Constraint requires a decomposition of a cognitive system into *components*. How is this crucial notion to be understood? According to Craver (2007) "[c]omponents are the entities in a mechanism – what are commonly called "parts." (128). He goes on to characterize the relationship between mechanisms and their components as follows:

> Organization is the inter-level relation between a mechanism as a whole and its components. Lower-level components are made up into higher-level components by organizing them spatially, temporally, and actively into something greater than a mere sum of the parts. (2007, 189)

It follows that an entity cannot be a component of *itself*. Moreover, the components (parts) of the system should be distinct structures, or at least, not characterized simply in functional terms, on pain of trivializing the mechanists' requirement on genuinely explanatory accounts in neuroscience.[7] As Piccinini and Craver (2011) note, the point of mechanistic explanation is to reveal the causal structure of a system; this requires that the components (parts) over which causal transitions are defined be understood as structures of some sort.

---

[7] Milkowski (2013, 55) argues that a mechanistic analysis must 'bottom out' in the *constitutive* level, the level at which "the structures that realize the computation are described."

Let's consider the two parts of the 3M Constraint in turn. Condition (a) requires that variables in the model correspond to components, activities, and organizational features of the target neural mechanism. In the case of FT models the variables range over the arguments and values of a mathematical function. But often, perhaps even typically, nothing in the function-theoretically characterized system corresponds to (states of) components of neural mechanisms.[8] The Marrian filter computes a function from intensity values at points in the image to the rate of intensity change over the image. The Ullman structure-from-motion system calculates the unique rigid structure compatible with 3 distinct views of four non-coplanar points. The presumption, of course, is that these mathematically characterized systems are *realized* in neural hardware, but in neither case is the implementing hardware specified at the level of its components parts and their organization. Shadmehr and Wise (2005) hypothesize that the motor control system that computes vector subtraction is realized in a network in the pre-motor cortex, but again, nothing in the FT characterization corresponds to (states of) *components* of the network and their organization.[9] And Sueng et al's (1996,

---

[8] Since the relevant variables in FT models are mathematical, the 3M constraint should be interpreted as requiring a mapping from values of the variables to *states* of component parts. This is how I will understand the constraint.

[9] I am not denying that computational theorists sometimes attempt to specify correspondences between variables in their models and (states of) components of neural mechanisms. In explaining *forward kinematics* – the computation of target location in body-centered coordinates from information about eye orientation and retinotopic location of target – Shadmehr and Wise appeal to the Zipser and Anderson's (1988) three layer neural network model of gain modulation. Some nodes in the model's input layers represent (correspond to) eye orientation and others retinotopic location of target; output layers represent (correspond to) target location in body-centered coordinates. Zipser and Anderson hypothesized that neurons in area LIP and area 7A in the parietal cortex play the relevant computational roles. It is not implausible, then, to describe input and output layers in the Zipser-Anderson model as *component parts* of a neural mechanism. Interestingly, the Zipser-Anderson models fails to count as genuinely explanatory by Kaplan's lights. He says:

> The real limitations on the explanatory force of the Zipser–Andersen model is that it is difficult if not impossible to effect a mapping between those elements in the

2000) model of oculomotor control posits an internal integrator without specifying any correspondence between variables over which the computation is defined and (states of) components of implementing neural hardware. I discuss this example in more detail below.

Let us turn now to condition (b) of the 3M Constraint, which requires that the dependencies among the variables in the model correspond to causal relations among components of the target neural mechanism. In the case of FT models the dependencies among the variables ranging over the arguments and values of the specified function are, of course, mathematical. The presumption that systems characterized in FT terms are realized in neural hardware – as they must be if the FT model is to be true of the organism – amounts to the idea that there exists a mapping from physical state-types to the arguments and values of the specified mathematical function, such that causal state transitions among the physical states are interpreted as mathematical relations among the arguments and values of the function. A complete computational explanation of a cognitive capacity will specify this mapping. Con-

---

> model giving rise to gain-modulated hidden unit activity and the neural components in parietal cortex underlying gain-modulated responses (arguably, the core requirement imposed by 3M on explanatory mechanistic models). (2011, 365-6)

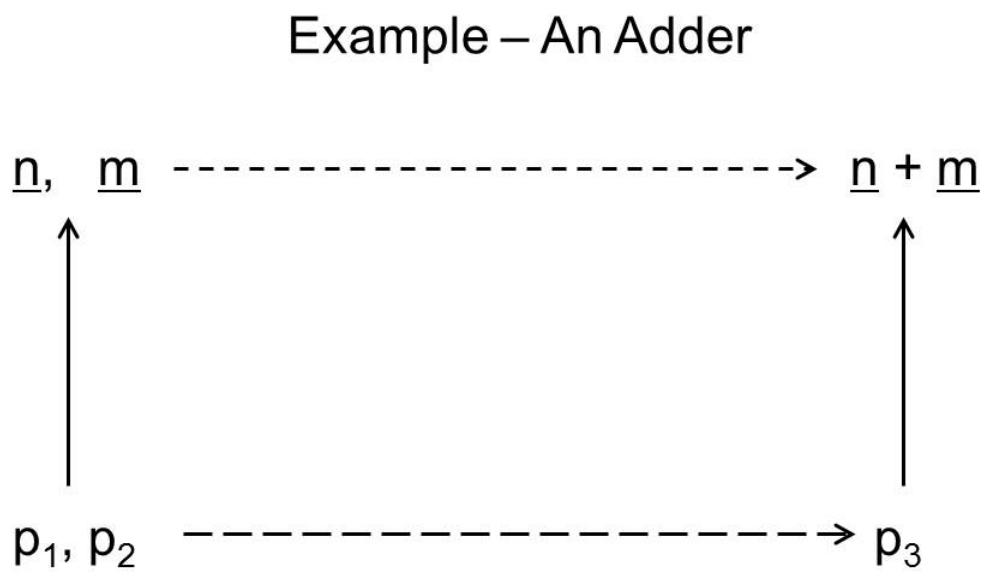
Kaplan cites two reasons why the model fails to be genuinely explanatory:

> First, it is difficult in general to effect a mapping between neural network models and target neural systems. There is typically only a loose and imprecise correspondence between network architecture and neural implementation (see, e.g., Crick 1989; Smolensky1988). (2011, 366)

Secondly, Kaplan notes that there are competing models of how gain modulation is implemented in the brain, each enjoying some empirical support, and so, he concludes, the Zipser-Anderson model is just a "how possibly" model and not genuinely explanatory.

An interesting question is whether, according to mechanists, the apparent failure of the Zipser-Anderson model to satisfy the 3M Constraint thereby undermines the explanatory *bona fides* of the Shadmehr-Wise function-theoretic model that it is supposed to implement. Presumably it does, since variables in the high-level characterization do not in fact correspond to components in an explanatory model of neural mechanisms. A consequence of the mechanist constraint is that any breakdown or lacunae in the decomposition (all the way down to basic physical constituents?) would seem to threaten the explanatory credentials of higher-level theories.

sider the following characterization of a device that computes the addition function (figure 1):



A physical system computes the addition function just in case there exists a mapping from physical state types to numbers, such that physical state types related by a causal state-transition relation ($(p_1, p_2) \rightarrow p_3$) are mapped to numbers n, m, and n+m related as addends and sums. Whenever the system goes into the physical state specified under the mapping as n, and then goes into the physical state specified under the mapping as m, it is caused to go into the physical state specified under the mapping as n+m.

It follows that the function-theoretic description provides an abstract characterization of causal relations among initial and end states of the realizing physical mechanism, whatever that happens to be. The physical states [$p_1 \ldots p_n$] that are characterized abstractly as the

arguments and values of the function (as addends and sums in the above example) in the complete computational model *may* count as components (in the sense explicated above) of the neural mechanism, but there is no reason to assume that they must. In precisely those cases where condition (a) of the 3M constraint fails to be satisfied – where the arguments and values of the specified function do not correspond to (states of) components of the neural mechanism – condition (b) will fail to be satisfied as well – the dependencies among the variables specified by the mathematical description will not correspond to causal relations among (states of) *components* of the target neural mechanism.

Seung's (1996, 2000) model of oculomotor memory illustrates this failure. It is worth examining the case in more detail.

The last 40 years has seen a good deal of experimental and theoretical work on oculomotor control.[10] Saccadic eye movements shift the eyes rapidly from one position in the visual field to another. Between saccades the eyes remain stationary; experimental results show that normal humans can hold their eyes still at arbitrary positions for twenty or more seconds, even in complete darkness. The most plausible explanation is that the brain maintains current eye position after a stimulus has gone by employing a short-term memory of eye positions. A wealth of experimental data supports the hypothesis of a multi-stable recurrent network located in the brainstem that takes as input transient eye velocities and gives as output persistent eye positions. It does so by accumulating input pulses, adding or subtracting (depending on the direction of movement) new inputs from the previous summation. In other words, it performs *integration*. A second, 'read-out' network reads the current position and stabilizes the eye by controlling the length-tension relationships of the muscles. Seung

[10] For general discussion see Robinson (1989), Glimcher 1999, and Leigh and Zee (2006).

(1998) describes the neural integrator as follows:

> In the oculomotor system, the integrator can be regarded as an internal model. The location of this internal model is known, unlike in other motor systems. As a steady eye position can be maintained without proprioceptive or visual feedback, the quality of the internal model is very good. And physiological studies of this internal model indicate that it is a recurrent neural network with a continuous attractor. (Seung 1998, 1257)

Encoding eye position in neural activity requires a continuous, analog-grade code, thus motivating the choice of a continuous attractor network. Figure 2 illustrates the continuous line attractor dynamics of the network.[11] A new stimulus changes the state of the network away from a line of fixed points. The network then settles on a new point along the attractor line; this point encodes the current eye position. Line attractor neural networks are posited to underlie a wide variety of motor control functions.[12]

---

[11] See Seung 1996, 1998. For general discussion of attractor networks see Amit (1989), Eliasmith and Anderson 2003, and Eliasmith 2005.
[12] Besides Seung 1996, 2000, see Shadmehr and Wise 2005.

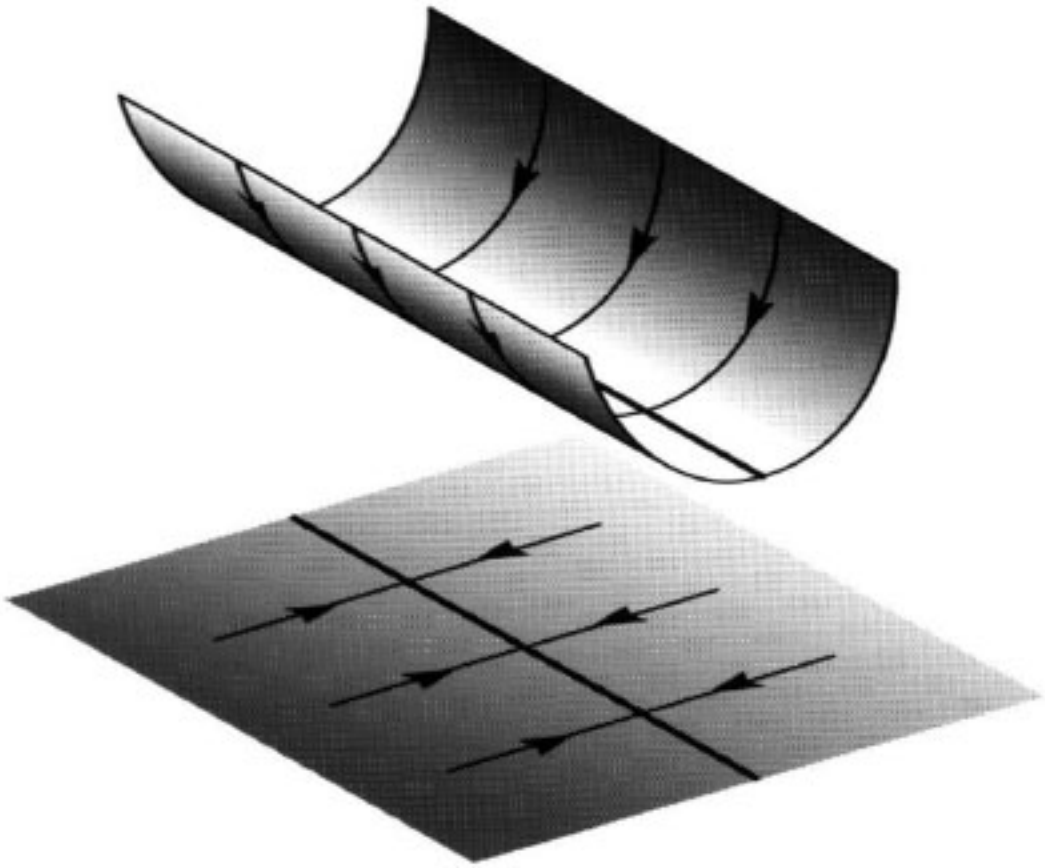


Figure (2) A state-space portrait for the eye-position memory network. All trajectories in the state space of the network flow toward a line attractor (thick line). Each point in the line is a persistent state that represents a different eye position. (From Seung 1996, p. 13340.)

Two features of the neural integrator are of special interest for present purposes. In the first place, the network has no proprietary input and output units; each unit is interconnected to all other units and can receive external stimuli (that is, pulse saccades). Secondly, no single unit or proper subset of total units represents an eye position; rather, only the total state of the network at a given time is a candidate for encoding a persistent

eye position. Points in the state-space portrait [figure 2] do not represent the activity of single cells, but rather the collective activity of the whole network. At any given moment the network "occupies" a point in the portrait, and it "aspires to" the line attractor. Points along the line attractor are collective states of the network that represent persistent eye positions.

Let's consider these two features of the network in light of the mechanists' 3M constraint. The computation effected by the integrator takes as arguments eye movement velocities and gives as values persistent eye positions. Condition (a) of the 3M constraint, recall, requires that the variables in the model "correspond to identifiable components, activities, and organizational features of the… mechanism that… underlies the phenomenon." (Kaplan 2011, 347) Arguably, neither variable of the FT model corresponds to (states of) components of the network that realizes the computation. With respect to the arguments of the function: as noted above, any of the networks' units can receive external stimuli from pulse saccades generating eye velocities. The mechanist may respond that eye velocities correspond to (are realized by) states of components of the mechanism; they correspond to states of *different* components for each run of the computation. But the mechanist cannot take this line for the *values* of the function. Persistent eye positions correspond to (are realized by) collective states of the whole network. On any plausible construal of 'component', collective states of the whole network do not count as components of the network. And while it is certainly true that the values of the integration do correspond to (are realized by) "activities" and "organizational features" of the network, a weakened construal of condition (a) of the 3M constraint that makes no mention of *components* and their interactions amounts to nothing more than the require-

ment that the model is realized by neural hardware; in other words, it imposes only the requirement that there exist a mapping of the sort depicted in figure 1.

Turning to condition (b) of the 3M constraint: since condition (a) is not satisfied – the arguments and values of the function do not correspond to (states of) components of the neural mechanism – condition (b) is not satisfied either. While the FT model gives an abstract characterization of causal relations between initial and end states of the attractor network, the dependencies among the variables specified by the FT description do not correspond to causal relations among *components* of the network. Since attractor networks of the sort that realizes the neural integrator are widespread in cognitive neuroscience, the 3M constraint is likely to fail for a wide variety of FT cognitive models. The mechanists' requirements on genuine explanation would have the unfortunate consequence of stripping much promising research in cognitive neuroscience of its explanatory interest.

To summarize the argument in this section: many FT models fail to satisfy the mechanists' strictures on genuine explanation. They do so for one of two reasons: either (1) there is a detailed and well-confirmed account of the neural mechanism that realizes the computation, but the relation between the FT description and the realizing mechanism is not of the specific type characterized by the 3M constraint, viz. a mapping from arguments and values of the computed function to (states of) components of the mechanism; or (2) the neural mechanism that realizes the computation is presently unknown, though theorists may have some idea of general features of the mechanism, such as where in the brain it must be located. In the second sort of case there is obviously much more theoretical work to be done. A computational model is not complete until the algorithm used to compute the function and the neural hardware that implements it have been specified. But there is no reason to think

that the neural realization has to be of the specific type favored by mechanists.

In the next section I discuss the specific features of FT models that make them explanatory, when they are. First, though, a more general point: claims can often be explanatory in the absence of realizing details. *That someone started a fire* can be an explanation of a forest fire. Of course, it is not a *complete* explanation; for that we would need to know about the chemical composition of the materials involved in the incident, the condition of the prevailing winds, and so on. But, as many have noted, explanation is typically *interest-relative*; sometimes the relevant interests are served without specifying the realizing details. The special sciences – including the sciences that purport to explain cognitive capacities – are continuous with ordinary practice in this respect.

4. *The explanatory credentials of function-theoretic models*

Computational models are proposed to explain our manifest success at some cognitive task – seeing 3D structure in the scene, understanding speech, grasping objects in nearby space, and so on. Justifying the FT description requires explaining how computing the value of the specified function contributes to the exercise of the cognitive capacity in question. The model is answerable, primarily, to behavioral data about the performance of the system in its normal environment. Thus, the theorist's first task is to characterize, in rough and ready terms, the organism's competence in the relevant cognitive domain – in what circumstances it is successful, and in what, perhaps rare, circumstances, it fails. Often the explanatory burden will be carried not by details of the realizing neural mechanism, about which very little may be known, but by features of the environmental context in which the mechanism normally operates. As noted above, Ullman's *structure-from-motion* mechanism is able to re-

cover the 3D structure of a moving object by computing the unique rigid interpretation consistent with three distinct views of four non-coplanar points only because in our world objects are typically rigid in translation. Appeal to a general environmental constraint (*rigidity* in this case) is crucial to the explanation of the organism's pattern of success and failure. Very little is known about the neural mechanism that implements the function (beyond the fact that areas V3/V3A and the dorsal parieto-occipital junction appear to be implicated).

As I have noted, an FT characterization of a cognitive mechanism resides at the topmost of Marr's explanatory levels, the so-called *theory of the computation*. It provides a canonical specification of the function computed by the mechanism, hence it answers a 'what-question': *what, precisely, does the device do*? But it also takes the first step in specifying *how* the system computes the cognitive function that is the explanandum of the theory: it computes the cognitive function, in its normal environment, *by* computing the specified mathematical function. However, the FT characterization does not "reveal the causal structure of the mechanism", as Kaplan (2011, 352) requires, except at a very high level of abstraction.

By its very nature an FT characterization is *multiply realizable* – it subsumes both natural and artifactual computers. Moreover, much of its explanatory 'oomph' depends on the fact that it is abstract; our grasp of a mathematical characterization – say a characterization of a system as an *adder* or an *integrator* -- is independent of any acquaintance we may have with particular (type or token) physical realizations of the mathematical description.

The idea that mental or other 'high-level' properties are multiply realized has recently come under attack. Opponents of multiple realizability argue that only the various structure-specific realizers of putative multiply realized properties count as genuine explanatory kinds.

The dialectical context of these arguments is an attack on non-reductive materialism. Reduction isn't to the point here, but if these arguments succeed in undermining the integrity of multiply realized kinds then the explanatory *bona fides* of FT models would be threatened.

Shapiro (2000) poses a dilemma: either the realizing kinds of a putative higher-level multiply realized property share many causally relevant properties or they don't. If the realizers share many causally relevant properties, then they are not distinct realizations. If they don't share many causally relevant properties, then any generalizations that apply to them will be "numbingly dull" (649). (Shapiro cites as examples of such numbingly dull generalizations that all realizers of mousetraps are used to catch mice, and that camera eyes and compound eyes both have the function of facilitating sight.) So either the higher-level kind is just not multiply realized or there is no motivation for subsuming the various distinct physical kinds under a higher-level (multiply realized) kind.[13]

FT kinds evade both horns of the dilemma. Shapiro says "… multiple realizations truly count as *multiple* realizations when they differ in causally relevant properties – when they make a difference to how they contribute to the capacity under investigation." (644) Corkscrews that differ only in color contribute in identical ways to removing corks, but hand calculators and human brains almost certainly differ in relevant causal powers, for example, in how they contribute to the system's capacity to compute the addition function. It is very likely that they employ different algorithms that require different realizing mechanisms. But the fact that these very different physical systems both compute the addition function – the fact that we can specify their behavior over a staggeringly large range of input conditions – is hardly "numbingly dull." So Shapiro's argument fails to show that FT kinds are not genuine-

[13] See Kim 1992 for a similar argument.

ly multiply realized.

Klein (2008) argues that there are no cases of genuinely multiply realized kinds in science. All putative examples either only support generalizations that are projectible *within* the restricted-realization kind, or, if they appear to support generalizations that are projectible across other realization kinds, turn out, on closer examination, to be non-actual *idealizations*, and so involve no ontological commitment to the higher-level kind. Materials science provides an example of the first sort of case: it classifies as *brittle* various materials – brittle steel and brittle glass, for example – that otherwise have very little in common. Of all that we know about brittleness in steel – that brittleness is proportional to hardness, that steel can be made less brittle by tempering, and so on – almost nothing applies to brittle glass. Discoveries about one realization-restricted kind of brittle material are not projectible to other realization-restricted kinds. Klein goes on to say:

> If there are [multiply realizable] kinds, they must be proper scientific kinds. If they are scientific kinds, then we should be able to project generalizations about them across all instances of that kind. But there aren't any such projectable discoveries; it looks like we must therefore abandon MR kinds – and not just in metallurgy, but in all of the special sciences, and psychology in particular. (2008, 162)

Klein concludes that scientific ontologies should include only "realization-restricted" kinds. The upshot is that FT kinds – which appear to subsume both biological subjects and artifacts, and hence are not realization-restricted kinds – should be eliminated; at best they are idealizations that don't literally apply to anything.

I will tackle the second disjunct of the dilemma first. FT models are not idealizations, in the sense that Klein has in mind. He says

> Idealizing models do not purport to describe the world… idealizing models are mere possibilia. Talk about them is false of anything in the actual world. You can't use them to predict anything…. Idealizations are typically used to explain the ceteris paribus laws that cover the (realization-restricted) kinds of particular special sciences. When scientists talk about the ideal gas, it is usually in the context of explaining the ceteris paribus laws that cover actual gasses. The ideally brittle solid is never cited on its own to explain anything…. (2008, 173)

To be sure, FT characterization does involve idealization. To describe a hand calculator as computing the addition function is to attribute to it a capacity defined on an infinite domain. The calculator's *actual* capacity is limited by the size of its display. A similar point applies to any biological system. And artifactual and biological computers are subject to various sorts of noise. They can fail to compute the specified function when they overheat or are exposed to harmful substances. Nonetheless, the FT characterization is intended to be *literally true* of the calculator, as is an FT characterization of a biological system in a computational psychological model. And, as I have noted, FT characterizations allow the prediction of the device's behavior across a wide range of input conditions, viz. those corresponding to the arguments of the specified function. So they are not idealizations in Klein's sense.

The argument against the first disjunct of Klein's dilemma – that empirical discoveries about one class of realizers do not project to other classes of realizers, and so commitment to a multiply-realized higher-level kind is not scientifically motivated – is somewhat less direct. It is true, of course, that what we know about the circuitry of the hand calculator is unlikely to be true of the brain (and vice versa). But the fact that lower level physical facts about one class of realizers are not projectible to other classes is beside the point. The under-

standing we gain of the capacities of a system (whether artifactual or biological) from FT models depends on the *abstract* character of the capacity attributed, not on any particular physical realization of that capacity. As I noted above, the explanatory payoff of FT characterization depends on the fact that the mathematical functions deployed in computational models – addition, integration, Laplacian of Gaussians, and so on – are well understood independently of their application in such models, and independently of our familiarity with computing devices, which of course is a relatively recent development. If this is right then Klein's case for the elimination of multiply realized kinds does not apply to FT kinds.

The upshot is that FT kinds are genuinely multiply realized; in fact they may be *sui generis* multiply realized kinds. They are not only *abstract*, but they are also *normative*. Theories of cognition are charged with explaining not just behavior, but, more importantly, cognitive *capacities* or *competences*, and FT models do so by positing further (mathematical) competences. In attributing a competence to a physical system – to add, to compute a displacement vector, and so on – FT models support attributions of *correctness* and *mistakes*. Just as the normal functioning of the system – correctly computing the specified mathematical function – explains the subject's success at a cognitive task in its normal environment, so a malfunction explains its occasional failure. Ingesting too much alcohol can cause neural systems to malfunction in any number of ways; one effect is that computational mechanisms may not compute their normal functions. 'The hand overshot the cup because the motor control system *miscalculated* the difference vector' is a perfectly good explanation of a motor control failure.[14]

This gives the FT characterization a kind of *autonomy* – the physical description that

[14] Of course, the hand may have overshot the cup for a variety of other reasons.

specifies the realizing neural mechanism does not allow the reconstruction of the normative notions of correctness and mistake.[15] The FT characterization explains the cognitive capacity by appeal to another competence not recoverable from the physical/causal details alone. But though the normativity inherent in the FT description cannot be accounted for at the level of realizing mechanisms there is nothing mysterious here. Look again at the adder depicted in figure 1. The bottom span of the figure specifies the physical state transitions that characterize the normal operation of the mechanism. When conditions are not normal – for example, when a human subject containing the neural adder is drunk, or a hand calculator is immersed in water – these physical state transitions may be disrupted. In other words, the system may be in the physical state(s) that (under the interpretation imposed by the mapping) realizes the arguments of the function, but fail to go into the physical state that (under interpretation) realizes the value of the function. The specification of physical state transitions (the bottom span of the figure) does not support attributions of correctness or mistake; the normative attributions are a consequence of the computational interpretation imposed by the mapping to the function.

In summary, the fact that FT characterizations are both *abstract* and *normative* (in the above sense) explains how FT models can be genuinely explanatory even absent an account of their neural realization.

5. *Objections and replies*

[15] This is not to deny that accounts of neural mechanisms may advert to such normative notions as *well-functioning* and *malfunction*. But physical/causal descriptions, even when they advert to functional notions, do not support attributions of *correctness* and *mistake*. They do not allow us to say that the mechanism *miscalculates* (or *misrepresents*) the vector from hand to cup.

I conclude by considering some objections.

Objection (1): FT models are not genuinely explanatory; they are what Craver (2006) and Kaplan (2011) call *phenomenological models (p-models)*. P-models "provide descriptions (often highly abstract, mathematical descriptions) of the phenomena for which explanations are sought… [but they] merely 'save the phenomena' to be explained." (Kaplan 2011, 349)

Reply: FT characterizations are not p-models – they do not just give a mathematical description of an observed regularity; rather they claim that the device *computes* a particular mathematical function. This distinction is important. The motion of the planets can be described (mathematically) by Kepler's laws, but the planets do not *compute* Kepler's laws, in the intended sense. The explanandum of a computational cognitive theory is a manifest cognitive capacity. An FT model is a hypothesis about how the system does it, by exercising a mathematical competence. The solar system has no manifest cognitive capacities that require appeal to mathematical competence. The objection that computational models, in the absence of realizing neural details, are just p-models, in other words that they are just descriptions of behavior, rests on a misconstrual of these models. The models give an abstract characterization of a mechanism that produces the phenomena by computing a mathematical function.[16]

[16] Putnam (1988) and Searle (1991) argue that every physical system computes every function. If every physical system does compute every function, in the sense at work in function-theoretic explanation, then the distinction between a system being merely describable mathematically and a system actually computing a mathematical function *collapses*, and computational models cannot be genuinely explanatory. The Putnam/Searle arguments have been widely discussed. For recent responses see Chalmers (2012), Egan (2012), and the other papers in *The Journal of Cognitive Science*, Vol. 12.

Objection (2): FT models do not describe "the real components, activities, and organizational features of the mechanism that in fact produces the phenomena." (Craver 2006, 361). They are mere 'how-possibly' models, rather than 'how-actually' models. As Kaplan puts it "the cost imposed by departing from [the mechanists'] view… is the loss of a clear distinction between computational models that genuinely explain how a given phenomenon is actually produced versus those that merely describe how it might possibly be produced." (2011, 346)

Reply: I doubt that that there *is* a sharp distinction between how-actually and how-possibly models. Weiskopf (2011) argues that the distinction is *epistemological*. As a model that purports to explain some phenomenon is more fully developed and acquires empirical support it can cross the threshold from 'how-possibly' to 'how-actually', though the latter verdict is always defeasible. FT models are hypotheses about how a system in fact exercises a particular cognitive capacity. A well-confirmed account of the algorithm used to compute the specified function and the neural structure that realizes the computation would, of course, increase our confidence that the model describes how the brain actually does it.

It should also be emphasized that FT characterizations in the first instance are specifications of the function that the device in fact computes; they are, one might say, *what*-explanations. Sometimes the characterization takes the form of a specification of an algorithm, in other words, an intensional specification of the function computed. In such cases, the algorithm is offered, not as an account of how possibly the device computes the function, but of what it computes and how *in fact* it computes it, as is evidenced by the fact that theorists would change the hypothesized algorithm were evidence to become available showing that it is not the algorithm used by the device. Initial hypotheses regarding the functions

computed and the algorithms for computing these functions are selected from the computational theorist's toolbox, but that fact does not undermine the claim that these are hypotheses about what the device actually does and how it actually does it. How else are theorists supposed to develop theories except by using what they know? These initial hypotheses are sometimes modified in response to input/output data to the point that eventually the device is said to compute a function that may be sufficiently different from the initial hypothesis that it can only be described in intensional terms.[17]

Objection (3): You say that FT characterization is *autonomous*. Isn't this just an expression of what Piccinini (2006) has called *computational chauvinism*, the idea that (as Johnson-Laird 1983, 9) put it: "[t]he mind can be studied independently from the brain… [that] psychology (the study of the programs) can be pursued independently from neurophysiology (the study of the machine and the machine code)."

Reply: No, it isn't computational chauvinism. I have explained the sense in which the FT level is autonomous – it characterizes the physical system in abstract terms, as a member of a well-understood class of mathematical devices. Moreover, the FT characterization is *normative* in a particular sense – it supports attributions of *correctness* and *mistakes*, notions not available at the level of realizing neural mechanisms. I am *not* claiming that we can fully explain cognition without understanding these neural mechanisms; in fact, I in-

[17] The computational model of natural language processing presented by Marcus (1980) is a case in point: the proposed model is an argumentation of a standard LR(k) parser of the sort that one might encounter in a graduate level course in parsing theory. The augmentations are dictated by observed features of human linguistic competence, and the resulting model can be characterized only intensionally.

sist that we cannot. The full explanation requires both an account of the realizing mechanism (though, as I have argued, the relation between the FT characterization and the realization may not satisfy the 3M constraint) and, typically, an account of the environment in which the cognitive capacity is exercised. The point is rather that 'the hand overshot the cup because the system miscalculated the difference vector' enjoys a sort of *explanatory* autonomy from the realization details, in the same sense that 'the forest fire was started deliberately' is explanatory, independent of details about the materials involved.[18]

References

Amit, D. J. (1989), *Modeling brain function: The world of attractor neural networks*. New York, NY: Cambridge University Press.

Bechtel, W. (2008). *Mental mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. London: Routledge.

Bechtel, W., & Abrahamsen, A. (2005), "Explanation: A Mechanistic Alternative," *Studies in History and Philosophy of the Biological and Biomedical Sciences 36*: 421–441.

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

Cannon, S. and Robinson, D. (1985), "An Improved Neural-Network Model for the Neural Integrator of the Oculomotor System: More Realistic Neuron Behavior," *Biological Cybernetics* 53: 93-108.

Chalmers, D. (2012), "A Computational Foundation for the Study of Cognition." *The Journal of Cognitive Science* 12: 323-357.

Craver, C. (2006), "When Mechanistic Models Explain," *Synthese* 153: 355–376.

Craver, C. (2007), *Explaining the brain*, Oxford: Oxford University Press.

---

[18] Thanks to Robert Matthews, Oron Shagrir, and participants at the Philosophy and the Brain workshop at the Institute for Advanced Studies at Hebrew University of Jerusalem in May 2013 for helpful comments on an earlier version of this paper.

Crick, F. (1989), "The Recent Excitement about Neural Networks,"*Nature 337*: 129–132.

Cummins, R. (1975), "Functional Analysis," *Journal of Philosophy* 72: 741–765.

Egan, F. (2012), "Metaphysics and Computational Cognitive Science: Let's Not Let the Tail Way the Dog," *Journal of Cognitive Science* 13: 39-49.

Egan, F. (forthcoming), "How to Think about Mental Content," *Philosophical Studies*.

Eliasmith, C. (2005), "A unified approach to building and controlling spiking attractor networks," *Neural Computation* 17(6): 1276-1314.

Eliasmith, C. and Anderson, C. (2003), *Neural engineering: Computation, Representation, and Dynamics in Neurobiological Systems*. Cambridge, MA: MIT Press.

Fukushima, K., Kaneko, C.R.S., and Fuchs, F.R. (1992). "The neuronal substrate of integration in the oculomotor system," *Progress in Neurobiology 39*, 609–639.

Gallistel, C.R. (1990), *The Organization of Learning*, Cambridge, MA: MIT Press.

Glimcher, P. W. (1999), 'Oculomotor Control', in R. A. Wilson and F. C. Kiel (eds), MIT Encyclopedia of Cognitive Science, Cambridge, MA: MIT Press, pp. 618–20.

Johnson-Laird, P. (1983), *Mental models: Towards a Cognitive Science of Language, Inference and Consciousness*. New York: Cambridge University Press.

Kaplan, D.M. (2011), "Explanation and Description in Computational Neuroscience," *Synthese* 183(3): 339-373.

Kaplan, D. and Craver, C. (2011), "The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective," *Philosophy of Science,* 78(4): 601-627

Kim, J. (1992), "Multiple Realization and the Metaphysics of Reduction," *Philosophy and Phenomenological Research*, 52: 1–26.

Klein, C. (2008), "An Ideal Solution to Disputes About Multiply Realized Kinds," *Philosophical Studies*, 140: 161–177.

Levy, A. (2013), "Three Kinds of New Mechanism," *Biology and Philosophy* 28: 99-114.

Leigh, R. J. and Zee, D. S. (2006), *The Neurology of Eye Movements (4th edition),* New York: Oxford University Press.

Machamer, P., Darden, L., and Craver, C. (2000), "Thinking about Mechanisms," *Philosophy of Science* 67: 1–25.

Marcus, M. (1980), *A Theory of Syntactic Recognition for Natural Language*, Cambridge, MA: MIT Press.

Marr, D. (1982), *Vision*, San Francisco: W.H. Freeman.

Milkowski, M. (2013), *Explaining the Computational Mind*, Cambridge, MA: MIT Press.

Putnam, H. (1988), *Representation and Reality*, Cambridge, MA: MIT Press.

Piccinini, G. (2006), "Computational Explanation in Neuroscience," *Synthese* 153: 343-53.

Piccinini, G. and Craver, C. (2011), "Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches," *Synthese* 183(3): 283-311.

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

Searle, J. (1993), *The Rediscovery of the Mind*, Cambridge, MA: MIT Press.

Seung, S. H. (1996), 'How the Brain Keeps the Eyes Still', Proceedings of the National Academy of Science USA 93: 13339–44.

Seung, S. H. (1998), "Continuous Attractors and Oculomotor Control," *Neural Networks* 11: 1253-8.

Seung, S. H., Lee, D. D., Reis, B. Y., and Tank, D. W. (2000), "Stability of the Memory of Eye Position in a Recurrent Network of Conductance-Based Model Neurons" *Neuron* 26: 259–71.

Shadmehr, R. and Wise, S. (2005), *The Neurobiology of Reaching and Pointing: A Foundation for Motor Learning*, Cambridge, MA: MIT Press.

Shapiro, L. (2000), "Multiple Realizations," *Journal of Philosophy* 97: 635–654.

Smolensky, P. (1988), "On the Proper Treatment of Connectionism," *Behavioral and Brain Sciences* 11:1–23.

Ullman, S. (1979), *The Interpretation of Visual Motion*, Cambridge, MA: MIT Press.

Weiskopf, D. (2011), "Models and Mechanisms in Psychological Explanation," *Synthese* 183(3): 313-338.

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