EXPLANATION IN COMPUTATIONAL NEUROSCIENCE: CAUSAL AND NON-CAUSAL

M. CHIRIMUUTA*

February 3, 2016

CONTENTS

ABSTRACT

This paper examines three candidate cases of non-causal explanation in computational neuroscience. I argue that there are instances of *efficient coding explanation* which are strongly analogous to examples of non-causal explanation in physics and biology, as presented by Batterman (2002), Woodward (2003) and Lange (2013). By integrating Lange's and Woodward's accounts I offer a new way to elucidate the distinction between causal and non-causal explanation, and to address concerns about the explanatory sufficiency of non-mechanistic models in neuroscience. I also use this framework to shed light on the dispute over the interpretation of dynamical models of the brain.

* mac289@pitt.edu. Penultimate Version. Forthcoming in BJPS.

1 INTRODUCTION

In recent philosophy of neuroscience and cognitive science there has been an overriding emphasis on mechanisms and mechanistic explanation[1] and some authors have gone so far as to present the mechanistic approach as the only game in town when it comes to explaining various phenomena—from sub-cellular signalling to computation in neural circuits and person-level decision making.[2] Following Woodward (forthcoming) I treat mechanistic explanation as a sub-type of causal explanation characterized by features such as reliance on decompositional methods and sensitivity to the details of implementation. A working assumption of the mechanist brand of philosophy of neuroscience has been that to explain a phenomenon is to describe the parts of the causal nexus that give rise to it.

Not surprisingly, a number of authors have detected a greater degree of methodological and explanatory pluralism amongst the sciences of the mind-brain.[3] On a number of occasions it has been argued that applications of dynamical systems theory (DST) yield non-mechanistic explanatory models.[4] In particular, Ross (2015) has argued that the explanatory patterns employed in highly abstract, dynamical models of spiking neurons (Izhikevich, 2010) are of the same form as the "minimal model explanations" first described by Batterman (2002) in the context of statistical mechanics. The published position of the mechanists has been that dynamical models are either not explanatory but merely predictive and descriptive (Kaplan and Craver, 2011), or that when they do offer explanations these are of a mechanistic sort [Bechtel (2011), Kaplan and Bechtel (2011)]. The debate over DST in neuroscience is far from settled. Yet if one does accept that dynamical models exemplify minimal model explanation, this raises the interesting question of whether such explanations are also *non-causal*. Robert Batterman and Collin Rice have argued that minimal model explanations in physics and biology are non-causal,[5] whereas Silberstein and Chemero (2013) hold that DST offers neuroscientists and cognitive scientists a distinctive genre of causal explanation.[6]

My central claim in this paper is that there are indeed instances of non-causal explanation to be found within neuroscience. I aim to convince you that the cases in neuroscience are at least strong as the familiar examples from physics and biology. Along the way, I will elucidate the distinction between causal and non-causal explanation by developing James Woodward's proposal that non-causal explanation occurs when we are able to answer questions about non-actual scenarios ("what-if-things-had-been-different" or "w-" questions), even though those scenarios are not the consequence of hypothetical interventions.

In Sections 2 and 3, I will focus on "efficient coding explanation" in computational neuroscience where, I will argue, one finds numerous

1 See e.g. Machamer et al. (2000), Craver (2007), Bechtel (2008), Kaplan and Craver (2011), Kaplan (2011), Kaplan and Bechtel (2011), Levy (2014) amongst many others.
2 Kaplan (2011); Piccinini and Craver (2011); Piccinini and Bahar (2013).
3 E.g. Weiskopf (2011), Barberis (2013), Serban (2015).
4 See e.g. Chemero and Silberstein (2008), Stepp et al. (2011), Silberstein and Chemero (2013).
5 Batterman (2010); Rice (2012); Batterman and Rice (2014)
6 An equivalent debate is ongoing over the status of network models in neuroscience. Craver (2014) argues, contra Huneman (2010), that examples put forward of *network* models providing non-causal, *topological* explanations of neural function are either mechanistic or non-explanatory. Due to limitations of space I will not discuss network models in this paper. However, it would be an interesting exercise to apply the framework I develop below to these kinds of examples.

instances of "distinctively mathematical", non-causal explanation, of the sort discussed by Lange (2013). Importantly, these explanations meet the mechanists' own criterion for explanatory sufficiency when they are able to answer w-questions (Kaplan, 2011, 354). In Section 4 I will turn to one recent example of a dynamical model of a region of the monkey brain. Here, it is less clear that the model offers a non-causal explanation and I will argue that the issue turns on whether or not one is willing to give a realist interpretation of the model components resulting from dimensionality reduction analysis employed by the model builders. In the remainder of this section I will say more about the relevant background, defining efficient coding explanation and presenting my preferred account of non-causal explanation.

1.1 Efficient Coding Explanation in Computational Neuroscience

In a recent publication I argue that models in computational neuroscience often yield a distinct, non-mechanistic, pattern of explanation which I call "efficient coding explanation" (Chirimuuta, 2014). The term "computational neuroscience" labels a broad research area which uses applied mathematics and computer science to analyze and simulate neural systems. That paper responds to the work of Kaplan (2011), which attempts to incorporate all explanatory models of this field within the mechanistic framework. The case turns on the particular example of the Gabor model of V1 receptive fields, where a mechanistic criterion for explanatory success, the "models to mechanism mapping constraint" (3M) [Kaplan (2011, 347), Kaplan and Craver (2011, 611)] fails, and yet the model is still able to provide counterfactual information, thus answering "w-questions" (Woodward, 2003).

How can this be?[7] Well, the models in question ignore biophysical specifics in order to describe the information processing capacity of a neuron or neuronal population. They figure in computational or information-theoretic explanations of why the neurons should behave in ways described by the model. So while, on the one hand, such receptive field models may simply be thought of as phenomenological descriptions which compactly summarise observed responses of neurons in primary visual cortex (Kaplan, 2011, 358 ff.), on the other hand, by analysis of the information theoretic properties of the Gabor function itself, one gains an explanation of why neurons with the properties captured by the model appear at this particular stage of visual processing.

In short, such models figure prominently in explanations of *why* a particular neural system exhibits a characteristic behaviour. Neuroscientists formulate hypotheses as to the behaviour's role in a specific information-processing task, and then show that the observed behaviour conforms to (or is consistent with) a theoretically derived prediction about how that information could efficiently be transmitted or encoded in the system, given limited energy resources. Typically, such explanations appeal to coding principles like *redundancy reduction* (Barlow, 1961)—the notion that more information can be transmitted through a cable (e.g. axon) of fixed bandwidth if some of the correlations between signals are removed. They do not involve decomposition of biophysical mechanisms thought to underlie the behaviour in question; rather, they take an observed behaviour and formulate an explanatory hypothesis about its functional utility.

7 See Chirimuuta (2014, section 5.1) for more detailed discussion of the points covered here.

It is worth saying a word about the notion of "efficiency" in play here. A feature of this research is that neuroscientists draw on knowledge of man-made computational systems and attempt to "reverse-engineer" the brain, looking for the "principles of neural design" (Sterling and Laughlin, 2015). A basic fact is that information processing makes substantial demands on resources, both in terms of the material required to build a computer or nervous system, and the energetic cost of computational processing. It is assumed, reasonably, that the explanation of many features of neural systems can be derived from consideration of resource constraints—the need to achieve good computational performance in spite of a relatively small resource budget. As Sarpeshkar (1998, 1602) writes:

> The three physical resources that a machine uses to perform its computation are time, space, and energy. Computer scientists have traditionally treated energy as a free resource and have focused mostly on time ...and space .... However, energy cannot be treated as a free resource when we are interested in systems of vast complexity, such as the brain. ...Energy has clearly been an extremely important resource in natural evolution. [Cf. Attwell and Laughlin (2001); Sterling and Laughlin (2015)]

So while, in what follows, it is instructive to compare efficient coding explanations to optimality explanations in biology—because both are in the business of comparing actual biological systems to theoretically optimal solutions—the field does not rely on the strong adaptationist assumption that the brain of humans, or any other animal, *is* somehow optimal.[8] One must instead make the weaker assumption that there is some process or mechanism—either evolutionary, developmental or occurring during the life of the organism as adaptation through neural plasticity—which causes the system to tend towards the optimal solution. Efficient coding explanations typically proceed without specifying what that process is.

1.2 Defining Non-causal Explanation

Efficient coding explanations often make fine-grained predictions about what would occur in counterfactual scenarios. This is possible because of the way that the efficiency of a computational procedure is sensitive to the nature of the particular task at hand. For example, neurons with a certain kind of receptive field structure might be the most efficient means to encode sensory information in one kind of environment, but not for another. Thus one can show how neural properties are counterfactually dependent on the evolutionary or developmental environment. For this reason I have argued that this branch of computational neuroscience employs a proprietary kind of non-mechanistic, causal explanation. Yet if one is willing to extend the notion of a mechanism to include the whole apparatus of natural selection and ontogenesis one might propose that computational neuroscience is still just in the business of discovering mechanisms. However, the assimilatory impulses of even the most flexible-minded mechanist would have to stop at

8 Nor do optimality explanations in biology always rest on this assumption (Godfrey-Smith, 2001). Instead, the point is to show that an observed feature has similarities with a theoretically predicted optimum, though there may be substantial departures from optimality due to structural or other constraints. Also, Wouters (2007) gives an account of non-causal "design explanation" in biology which does not depend on any claims about the optimality, or near optimality, of biological systems.

the idea of distinctively mathematical non-causal—and non-constitutive—explanation. So a central motivation for exploring the question of whether non-causal explanations occur in neuroscience is as a way to get clearer on the limits of the mechanist framework.

Woodward's interventionist theory of causal explanation has been extremely influential in the philosophy of neuroscience, and Woodward's proposal that explanatory sufficiency is tracked by the ability of a theory or model to address w-questions is accepted by authors such as Kaplan and Craver. The basic intuition is that, "a successful explanation should identify conditions that are explanatorily or causally relevant to the explanandum: the relevant factors are just those that 'make a difference' to the explanandum in the sense that changes in these factors lead to changes in the explanandum" Woodward (forthcoming, 5).

Interestingly, Woodward (2003, 221) suggests that the ability to address w-questions may range beyond causal explanation, writing:

> the common element in many forms of explanation, both causal and noncausal, is that they must answer what-if-things-had-been-different questions. When a theory tells us how Y would change under interventions on X, we have (or have the material for constructing) a causal explanation. When a theory or derivation answers a what-if-things-had-been-different question but we cannot interpret this as an answer to a question about what would happen under an intervention, we may have a noncausal explanation of some sort.

Woodward gives the example of the hypothesis that the stability of the planets is counterfactually dependent on the four dimensional structure of space-time. What if space-time had been six-dimensional? There is no intervention associated with this question;[9] but the hypothesis is that if things *had* been different then planetary orbits would indeed be less stable.

In this paper I follow Woodward in defining an "intervention" as an idealized, unconfounded experimental manipulation of one variable which causally affects a second variable only via the causal path running between these two variables (Woodward, 2013, 46).[10] Like various other authors,[11] I believe it is useful to de-couple the counterfactualist parts of Woodward's account of explanation from the causal, interventionist ones and thereby develop an account of non-causal explanation. Moreover, I propose that this account be integrated with Lange's notion of "distinctively mathematical explanation" to give a clearer standard for differentiating causal from non-causal explanations than is often employed in the literature.[12]

9 For one thing, if God in a new act of creation were to change the dimensionality of space-time, this could not be thought of as a possible intervention to be performed by finite beings. It would be contentious to extend an interventionist account of causation to acts of creation by infinite beings. More to the point, the theory of general relativity tells us that the counterfactual dependence of planetary stability on the geometry of space-time is not a result of any causal relationship between these two. For this reason, we cannot think of alterations in space-time which result in changes in planetary stability as interventions in Woodward's sense. See definition at the start of next paragraph, and see Woodward (2014, 702).

10 I take this to be uncontentious since many mechanist authors have adopted Woodward's interventionist approach to causation, most notably Craver (2007).

11 See Bokulich (2008), Bokulich (2011), Saatsi and Pexton (2013).

12 A new paper by Baron et al. (forthcoming) independently hits upon this idea of using the framework of counterfactual explanation to characterise distinctively mathematical explanation. Their more formal presentation of this synthesis is a useful supplement to the examples I present below.

Marc Lange analyses distinctively mathematical explanations of regularities and events in terms of the modal strength of mathematical facts, in comparison to ordinary causal laws. For example, Lange (2013, 488) writes:

> That Mother has three children and twenty-three strawberries, and that twenty-three cannot be divided evenly by three, explains why Mother failed when she tried a moment ago to distribute her strawberries evenly among her children without cutting any.

This is conceptually different from any causal explanation that mentions, for instance, that the attempt made Mother hungry and frustrated ('hangry') and so she ended up eating two of the strawberries herself, or that one of the little darlings stole from the other, etc.

Thus I share Lange's view that distinctively mathematical explanations employed in science are non-causal ones (Lange, 2013, 506). Furthermore, I propose that we bring Lange's notion of modal strength in non-causal explanation to bear on Woodward's idea that non-causal explanations occur when we show that there are dependencies between the explananda and explans which cannot be understood in interventionist terms. In such cases, knowledge of the dependencies does not show us how things would be in a range of counterfactual scenarios in which we perform manipulations on the explananda. Instead, we are told how things would be under certain *impossible* scenarios in which the laws of mathematics are altered.[13] This of course assumes that *counterpossible* statements—counterfactuals or subjunctive conditionals with impossible antecedents—can be non-vacuously true. I invite the reader to consider that it is intuitive that the statement, 'if thirteen were evenly divisible by three, then I could share my baker's dozen of doughnuts equally amongst my three best friends' is non-vacuously true, and that the statement 'if thirteen were evenly divisible by three, then a baker's dozen of doughnuts would be a healthy snack' is non-trivially false; what is more, there is a recent literature on the semantics of counterpossibles which underwrites these intuitions.[14]

In the examples I present in the following two sections, we have a non-causal explanation which is reliant on a trade-off demonstrated in the theory of information. Such trade-offs are candidates for being brute mathematical facts—nothing could be done to make them not obtain. For this reason, we can think of the trade-offs as modally strong mathematical facts, in Lange's sense, and as yielding information about counterfactual dependencies which go beyond interventionist interpretation, in Woodward's sense. Like Lange (2013), but unlike Saatsi and Pexton (2013), I intend my account of non-causal explanation in neuroscience to cover explanations both of particular events and regularities, since the kind of mathematical facts that my examples depend on do yield explanations of both sorts.

In contrast to the account of non-causal explanation just outlined, Robert Batterman and Collin Rice have focussed on the way that certain models in physics, biology and economics idealise away from the causal processes which lead up to a phenomenon. Such models result in representations of the causes of phenomena which are, at best "caricatures" (Batterman and Rice, 2014). So these authors argue that the relevant notion of explanation is not a causal one since the explanation of a phenomenon does not arise

13 While it is thought not even God could affect such interventions.

14 See e.g. (Brogaard and Salerno, 2013) and (Bjerring, 2014). I am very grateful to an anonymous reviewer for pointing me to this literature.

because we are provided with information about what caused it. This is how Rice (2015, 600) makes the case:

> When considered together, these idealizations entail that optimality models usually provide little, if any, accurate information about the actual causes, or causal mechanisms, within the model's target system(s). In the end, the highly idealized optimality model represents mathematical relationships between constraints, tradeoffs, and the system's equilibrium point that *do not mirror any causal relationships (or processes) in the target system*. Put different, optimality models fail the kind of "model-to-mechanism-mapping requirement" of causal theories.

This is a less stringent notion of non-causal explanation than the one that I will employ below. Although all of my examples of distinctively mathematical explanations do abstract away from causal details, this feature is not what makes them non-causal (Lange, 2013, 506). While my cases of efficient coding explanations involving trade-offs do bear comparison with Rice's discussion of optimality models in biology, I would like to emphasise an important difference, namely that the optimality models discussed by Rice (e.g. the Fisher model of the 1:1 sex ratio) posit trade-offs which *could* be subject to experimental intervention (e.g. through engineering of a species or environment such that the cost of producing sons is not equal to the cost of producing daughters), whereas the trade-offs central to my examples are in a stronger sense fixed. This is because they are mathematical rather than empirical facts.

Irvine (2014, 12) argues that if the equilibria discovered in optimality modelling show stability or robustness in the face of experimental interventions, then the model cannot offer causal explanation. Again, I suggest this is too loose a notion of non-causal explanation because she intends it too include models whose equilibria could shift if we were to change "something fundamental about the system" such as intervening on inheritance mechanisms or introducing random fluctuations in the environment. But these are just different kinds of interventions—impractical or technically infeasible—but still within the remit of causal analysis. Irvine's proposal does not give us a clear cut way of deciding whether an explanation is causal or non-causal. In the next two sections of the paper I will show that it pays to uncover the use of mathematical facts in efficient coding analyses because these offer cases of non-causal explanations in neuroscience which are analogous to Lange's examples from biology, and which can be incorporated into Woodward's counterfactualist framework.

2 CASE I: HYBRID COMPUTATION

My first example is an explanation of the efficiency of the brain—thought of as a biological computational system—as being due to the advantages of hybrid computation.[15] The account was presented by Rahul Sarpeshkar from the department of electrical engineering and computer science at MIT. Sarpeshkar's research aims to use insights gleaned from our understanding of computation in biological systems in order to build more powerful and

15 Needless to say, my arguments do not turn on whether or not this explanation is disputed or accepted within the scientific community; rather, it serves to illustrate a pattern of explanatory reasoning. Recent work on this topic is reviewed by Sterling and Laughlin (2015, chap. 10).

efficient artificial computers (Sarpeshkar, 2010). His analyses must often abstract away from the implementational details which obviously differ tremendously from biological to man-made devices, and they consider the problem of computation in more purely mathematical terms. At the same time, this research is in the tradition of other efforts to apply information theory to neural systems in order to gain understanding of why the nervous system has the anatomy and physiology that it does.[16]

Since the birth of modern computing and the concurrent rise of quantitative neurophysiology in the mid twentieth century, many have compared these two very different kinds of information processing devices and have pondered the question of whether computation in the brain is analogue or digital.[17] It is known that brains are vastly more energy efficient than any digital computer. One estimate for the power (energy consumption) of the human brain is 12W (Sarpeshkar, 1998, 1601), whereas a supercomputer such as IBM's 2007 'Roadrunner', engaged in an equivalent number of around 10^{15} processing events per second (i.e. 1 petaflops) runs at 2.4 MW (Komornicki et al., 2009). A question posed and discussed by various neuroscientists is how to account for the impressive energy efficiency of neural tissue. Sarpeshkar (1998) considers this question by examining the relative efficiencies of analogue and digital computational systems.[18]

The key characteristic of a digital system is that, by definition, one of its components (e.g. a wire) can only represent 1 bit of information at a given time. This is because signals are all or nothing events, so the wire has only two informationally relevant states—on or off, 1 or 0—which amounts to 1 bit of information.[19] In contrast, for an analogue system the signal varies continuously with some physical variable of its components (e.g. voltage), so any one component has the potential to represent countless bits of information at a given time (Sarpeshkar, 1998, 1605). For example, to transmit 4 bits of information, one would need four separate components (wires) if coding digitally, but the same amount of information could be transmitted on a single wire in an analogue system, so long as 16 different physical states of the wire (e.g. voltages) could be unambiguously associated with 16 different signals.[20] Note that these are not empirical claims about copper wires, axons, or any other bits of hardware; they are, if you like 'analytical' statements about what we mean by representing or transmitting information in these different ways.

It follows that analogue systems are far less hungry for resources—raw materials and energy needed to build and maintain the signalling

16 See Cover and Thomas (2006) for more on information theory and Rieke et al. (1999) on the application to neuroscience.

17 E.g. MacKay (1991, 40) quoted by Husbands and Holland (2008): "Later in the 1940s, when I was doing my Ph.D. work, there was much talk of the brain as a computer and of the early digital computers that were just making the headlines as 'electronic brains.' As an analogue computer man I felt strongly convinced that the brain, whatever it was, was not a digital computer. I didn't think it was an analogue computer either in the conventional sense." See also von Neumann (2000), McCulloch and Pitts (1943), MacKay and McCulloch (1952).

18 To be more precise, the important contrast is between analogue and *asynchronous digital* or *pulsatile* systems. These are systems in which the signals are discrete, all or nothing events, but they operate in a continuous time frame. Neuronal spikes are a good example. Computer microprocessors, in contrast, send their discrete signals on a strictly clocked time schedule, and this is what is normally meant by digital computation.

19 Where information is defined in the Shannon-Weaver sense as $\log_2(N)$, N being the number of possible and equally probable states of the wire.

20 Of course, in a true analogue system the signal is a continuous function of a physical magnitude such as voltage so the idea of associating just 16 different signals with 16 discrete voltages in the wire is artificial. But it serves when making the comparison in resource consumption across digital and analogue systems.

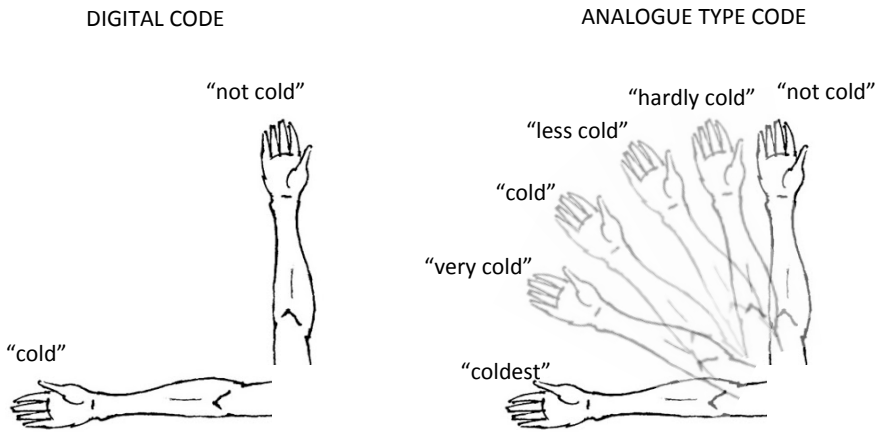

Figure 1: Illustration of the Trade-off between Resource Consumption and Noise Susceptibility. Given a fixed material resource which is subject to random fluctuations in its physical states, e.g. one person's arm, a digital code can only transmit 1 bit of information, signalling two different possible states, "cold" and "not cold" . In the analogue style code, the fixed material resource can be used to transmit an indefinite amount of information, since meaningful signals are defined across a continuous array of physical states. The angle of an arm from 0 to 90° is a physical continuum, and in this analogue style code, 6 states of the world are associated with 6 different ranges of arm angles. However, it is clear that the more states one tries to encode, the more susceptible the system is to noise. Here, any slight wobble of the arm could lead to an error in signal transmission.

components—than are digital ones. However, the downside of analogue computation is that such systems are much more susceptible to corruption due to noise—random fluctuations in the physical states of the components—than are digital ones. To take our example of a single wire used to transmit 4 bits of information, it is easy to see that if there are random changes in the voltage due to motion through magnetic fields, for example, then what started out as a 2V signal could end up being received as a 2.5V one. And the more information one attempts to transmit through a single wire, the greater the problem because the difference between the physical magnitudes encoding the signals must decrease. Again, this problem is not an empirical observation of the behaviour of metal wires but occurs for any signalling system whatever its material realisation, as Figure 1 illustrates.

Thus there is a rigid trade-off between economy of resource consumption and susceptibility to noise.[21] This trade-off is inherent to the definitions of noise, information and signal within the mathematical theory of information, and could not be altered through any empirical intervention. Indeed, noise is a result of the physical makeup of the components (e.g. random electron motion in nano wires, or random ion flux across neuronal membranes), and one may successfully engineer less noisy components. But the trade-off still obtains: the more information one tries to send through

21 In my exposition I concentrate on resource consumption in terms of number of material components required to transmit a given amount of information. The trade-off also occurs with respect to energy consumed to generate signalling states. One could increase the maximum voltage of a wire in order to create a greater 'spacing' between the voltages of signalling states, and thus make the system more resistant to noise corruption, but this would obviously require a bigger energy investment.

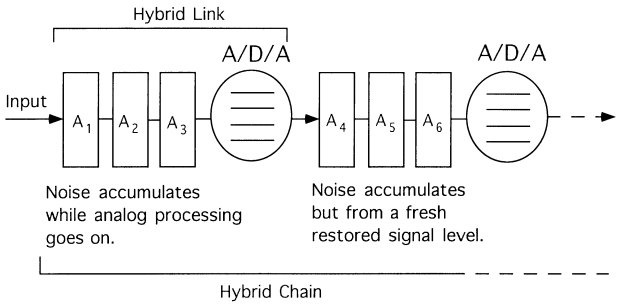

Figure 2: Hybrid Computation. Each "hybrid link" combines analogue processing with an analogue-digital-analogue (A/D/A) converter to clean up the signal. Sarpeshkar (1998, 1622) writes that "hybrid chains allow us to operate with the precision and complexity characteristic of digital systems, while doing efficient analog processing". Sarpeshkar (1998, fig. 5), *permission needed*.

one's components, the more the signal will be affected by noise. Given any set of components the more economically one tries to use those components, the more one's signals are going to be corrupted by noise.[22]

The point of Sarpeshkar's analysis is to sketch out an optimally efficient system, given this trade-off. The 1998 paper contains a number of different calculations to show, for example, the point at which a given analogue system becomes 'useless' due to noise accumulation. Sarpeshkar proposes *hybrid computation* as the best general solution to the problem of building a system which is resilient to noise (like a digital computer) but economical with resources (like an analogue computer). As Sarpeshkar (1998, 1636) puts it, "hybrid computation combines the best of the analog and digital worlds to create a world that is more efficient than either." The basic idea is to alternate between digital and analogue processing steps using digital-analogue converters. This way, one has the benefit of small chunks of efficient analogue computation, in which noise accumulates, interspersed with digital processing to 'clean up' the signal. The idea is illustrated in Figure 2.

Most of Sarpeshkar's analysis is illustrated in terms of metal wires and other artificial electrical components; but he is keen to argue that his conclusions generalise to biology (Sarpeshkar, 1998, 1634 and 1636). Furthermore, it is interesting to consider neuronal physiology in terms of the hybrid hypothesis:

> Action potentials are all-or-none discrete events that usually occur at or near the soma or axon hillock. In contrast, dendritic processing usually involves graded synaptic computation and graded nonlinear spatiotemporal processing. The inputs to the dendrites are caused by discrete events. Thus, in neuronal information processing, there is a constant alternation between spiking and nonspiking representations of information. . . . This alternation is reminiscent of the constant alternation between

22 As a historical aside it is worth noting that back in the 1950's John von Neumann, one of the founders of modern digital computing, observed that one could offset the effect of noisy components in the brain by using more of them in a code with built in redundancy, essentially highlighting this trade-off between the economy and reliability of a coding system (von Neumann, 2000).

> discrete and continuous representations of information in Figure 5 [Figure 2 above]. Thus, it is tempting to view a single neuron as a D/A/D.[23] (Sarpeshkar, 1998, 1630)

It is worth underlining some of the various claims that Sarpheshkar makes in the article. In the passage just quoted, he offers an interpretation of an often observed fact of neural physiology—the flow of electrical activity from dendrites to axons, and then on to dendrites of another neuron—as an implementation of hybrid computation. It could be said that he explains this fact by appealing to the efficiency of hybrid computation. But more central to his project is the explanation of the efficiency of the brain in its entirety—the fact that a man made super computer consumes orders of magnitude more energy than a biological brain. This fact is explained by showing that hybrid computation is the optimal solution to the problem of maximising resilience to noise while minimising resource investment, and by showing that it is plausible that hybrid computation is implemented in biological brains. As he states in the the abstract of the paper

> "Our results suggest that it is likely that the brain computes in a hybrid fashion and that an underappreciated and important reason for the efficiency of the human brain, which consumes only 12 W, is the hybrid and distributed nature of its architecture." (Sarpeshkar, 1998, 1601)

At the heart of Sarpeshkar's account is the demonstration of the costs and benefits of analogue and digital computation, and the optimality of the hybrid combination. Thus, I argue, the efficiency of the brain is explained, non-causally, by its implementation of hybrid computation. In support of this claim I will now compare Sarpeshkar's account with some standard examples of non-causal explanation.

My first point is that Sarpeshkar's proposal answer what-if-things-had-been-different questions. According to his discussion, the efficiency and reliability of the brain are counterfactually dependent on the efficiency and reliability of hybrid computation, which is in turn dependent on a trade-off between resource consumption and reliability. If this trade-off did not occur, then a purely digital or purely analogue system could in principle satisfy both of the desiderata, and the physiology of brains would have been very different. At the same time, any such scenario cannot be interpreted as the result of a possible intervention because the existence of the trade-off is not an empirical fact about the properties of actual objects but the result of our information theoretic definitions of analogue and digital coding schemes (see figure 1). That is to say, the point is not that some materials make more reliable signalling systems than others, but that given any physical system, and given a fixed resource investment, the amount of information transmitted per by the system trades off against the susceptibility of the signal to be corrupted from noise.

This is like Woodward's example of the non-causal explanation of the stability of planetary orbits, except that we are clearly talking about what would obtain *if the principles of information theory had been different*. In other words, the information theoretic explanation of the efficiency of the brain

23 Sarpeshkar (1998, 1630) adds these words of caution: "However, although the firing of a spike is a discrete event, it does not imply that it encodes information about a discrete state. The information encoded by a spike is meaningful only in relation to spikes in different neurons, or in relation to earlier or later spikes in the same neuron. If these relationships are analog, then all-or-none events do not imply the encoding of discrete states."

informs us about counterfactual (or counterpossible)[24] scenarios in which the laws of information theory are different. Tinkering with information theory and working out the implications for coding systems cannot be thought of as a causal intervention. One may rightly worry that such information is not as useful as counterfactual information about the effects of causal interventions.[25] Nevertheless, it is worth pointing out that all these kinds of explanations can be accommodated by Woodward's scheme for non-causal explanation.

We can also ask if this example falls within the framework of Lange's "distinctively mathematical explanation". To this end, it is worth comparing it with the famous bridges of Königsberg example. The explanation for why no-one has ever managed to cross all of the bridges in the city of Königsberg (in their 1735 configuration) just once, without ever doubling back over a bridge, or resorting to flight or swimming, depends on the characterisation of the set of bridges as a network of nodes and edges in graph theory. As it happens, each of the four nodes (rather than the requisite number of two nodes) is touched by an odd number of edges, and it is a modally strong mathematical fact that such a network cannot allow for the desired sequential crossings (Lange, 2013, 489). Likewise, one may interpret Sarpeshkar's analysis as showing that it is a modally strong mathematical fact that hybrid computation is the optimal way to satisfy the twin constraints of efficiency and robustness in the face of noise. So just as the graph theoretic properties of the bridge layout explain the observed limitations on crossing routes, the hybrid computational properties of the biological brain explain its efficiency.

It is also important to note that non-causal explanations of biological systems are focussed on the problem of specifying why particular strategies are highly efficient or theoretically optimal; that is, they have a specific kind of explanandum. We can, in turn, then explain why natural selection would have settled upon these strategies. But it is important to keep the causal explanation of 'how did this trait evolve?' separate from the non-causal explanation of why a certain trait may be optimal. For example, if one asks simply, 'why should the brain operate like a hybrid computer?', the response could be, 'because during the evolution of the brain there was strong selective pressure for the brain to be efficient and reliable, and hybrid computation is demonstrated theoretically to be the best way to satisfy the conflicting demands of efficiency and reliability.' Here, a causal and a non-causal explanation rub shoulders; but when arguing that distinctively mathematical, non-causal explanations take place it is important to be specific about the explanandum so that one can isolate the non-causal 'component' of broader explanations.

Lange makes a similar point in his discussion of the explanation of the hexagonal shape of honeycombs.

> The explanation is that it is selectively advantageous for honeybees to minimize the wax they use to build their combs—together with the mathematical fact that a hexagonal grid uses

24 Depending on how one understands the modal strength of information theory. It is conceivable that key concepts in information theory could have been defined differently and the theory still be consistent. So the Shannon-Weaver definitions do not have the obvious modal strength of "13 is a prime number". That said, the important point here is that any counterfactual dependency between the properties of coding schemes and information theory cannot be undertsood as a causal one, just as the dependency of planetary stability on the dimensionality of space-time is not causal.

25 This worry is the topic of another paper (see AUTHOR).

> the least total perimeter in dividing a planar region into regions of equal area .... This explanation works by describing the relevant features of the selection pressures that have historically been felt by honeybees, so it is an ordinary, causal explanation, not distinctively mathematical. But suppose we narrow the explanandum to the fact that in any scheme to divide their combs into regions of equal area, honeybees would use at least the amount of wax they would use in dividing their combs into hexagons of equal area.... This fact has a distinctively mathematical explanation. (Lange, 2013, 499-500)

However, in order to support the point that Sarpeshkar's explanation is distinctively mathematical, we would need to say more about the details of his quantitative analysis. The description above of the trade-off and the optimality of hybrid computation is based on Sarpeshkar's qualitative ("intuitive") account of these coding schemes. He also gives a mathematical description of the pro's and con's of analogue and digital computation, presented as a comparison between the "resource precision curves" for analogue and digital systems (Sarpeshkar, 1998, figure 3). What these show, on the one hand, is that the relationship between resource consumption and signal-to-noise ratio for digital systems is compressive, such that digital systems can increase precision by many orders of magnitude and only incur minor increases to an already high baseline level of cost. On the other hand, we see that for analogue computation there is a an expansive relationship between resource consumption and signal-to-noise ratio, and also that the maximum precision of the system is strictly bounded by thermal noise. When precision is low, costs are extremely low, but costs increase dramatically when the analogue system is operating in a more precise regime. At a certain signal-to-noise ratio ("the crossover point"), the costs of analogue computation begin to outpace those of digital computation.

The overall point is that analogue systems offer very cheap computation, but only if one is willing to tolerate a poor signal to noise ratio. The question now is, *why should we think of this as a mathematical fact?* One reason to think that the facts summarised in the resource precision curves are straightforwardly *empirical* is that the points plotted in such graphs are the outputs of equations which have empirically measurable parameters, such as the length and width of transistors. And as Sarpeshkar (1998, 1615) writes, "[t]he exact location of the crossover point will depend on the task, technology, and ingenuity of the analog and digital designers."

The reason why there is this contrast between cheap analogue computation and costly but precise digital computation boils down to the fact that the resource precision curves follow a logarithmic relationship for digital computation, while for analogue computation power consumed is proportional to signal-to-noise ratio, or to S/N^2, depending on the kinds of transistors used. These equations are derived from examples which cite specific parameter values, kinds of transistors, and make concrete assumptions about the noise distribution. The key question is whether the resource precision curves would take these forms given any parameter values and reasonable assumptions for analogue and digital systems, or if the results could be qualitatively different with different values and assumptions. I suspect that the qualitative form of the results would be unchanged but since I am not in a position to offer a proof of this assertion, I concede that we have an open question about whether Sarpeshkar's

explanation is distinctively mathematical in this strict sense. In the next section I present an example which is more clear cut on this issue.

Another potential worry here is that the existence of the trade-off illustrated in figure 1, and which I have argued is a result of our information theoretic definitions of the coding schemes, is due instead to the empirical fact that the states of all physical systems are liable to random fluctuations and hence introduce noise. In a noiseless world, we would assume, analogue computers would outperform hybrid and digital ones. Moreover, the objection goes, the trade-off would not be there in the noiseless world and so its existence is not a fact about information transmission that goes beyond any empirical facts. In response I urge that even if the laws of nature were altered such that we have perfectly noiseless physical systems for building brains and computers, we should think of the trade-off itself as being there—even though natural selection and human engineers would not be aware of it—because of they ways that noise and information are defined theoretically. Consider an analogous scenario: in contemporary society time spent at leisure trades off against time spent at work; the optimal solution to this trade-off is known as the 'work-life balance'. In a future Utopia where robots do all the work, nobody needs to spend time earning a wage so everyone spends all their time at leisure. People will not even be aware of the trade-off. But that does not alter the fact that by our definitions, time spent at leisure is not time spent earning a wage, and vice versa. In some sense the trade-off is still there, but empirical conditions do not make it apparent.

That said, it must be admitted that the empirical fact that actual physical systems are noisy is bearing substantial explanatory weight here—the trade-off is only relevant to brain physiology and computer design because of that fact. More generally, the notion of a trade-off only makes sense if we assume the presence of certain constraints. So in the final section of this paper I will consider the question of whether Sarpeshkar's explanation is ultimately a causal one because it must refer to such empirical facts. Before moving on, it is worth noting that nothing turns on whether one uses Sarpeshkar's analysis to explain a feature of one particular brain or a trait of brains in general. Thus, I would argue, this kind of non-causal explanation is applicable to singular states of affairs as well as regularities.

3 CASE II: THE GABOR MODEL REVISITED

In the previous case we can observe a close proximity between causal, evolutionary and non-causal, mathematical explanations in biology and neuroscience. We see this again when we compare different efficient coding explanations of the response properties of neurons in primary visual cortex (V1). As discussed in detail in Chirimuuta (2014, §5.2), the 2D Gabor function (product of a sinusoid with a Gaussian envelope) is part of the "standard model" of V1. It has long been observed that neurons in this area respond selectively to the visual presentation of short, bar like stimuli of a specific width and orientation. The parameters of the Gabor function can be adjusted to capture this selectivity pattern (the *receptive field*, RF), and this raises the question of why V1 neurons have receptive fields that can be fit by this equation.

I previously argued that the Gabor function is central to causal, but non-mechanistic explanations of V1 response properties. Here, I contend that

there is also a non-causal part of the picture. When Gabor (1946) first introduced the 1D version of the function it was in the context of signal engineering. Given the problem of analysing information contained in signal waves, he noted that there is a trade-off between one's ability to accurately decode temporal and spectral (frequency) information, which is known as "Heisenberg-Weyl uncertainty". The more reliably one decodes the arrival time of a signal, the less reliably one decodes its frequency, and vice versa. One cannot, at the same time, be maximally certain about both of these parameters. For example, Fourier analysis provides a very accurate analysis of the spectral composition of a wave, but temporal information is lost. Gabor proved that his function, which in effect performs a temporally localised Fourier analysis, provides the optimal balance between recovering temporal and spectral information. This, we can say, is a modally strong mathematical fact.

In one of the first papers to model V1 responses with the Gabor function, Daugman (1985, 1160) pointed out that the Gabor-like properties of V1 receptive fields could be the biological solution to the problem of jointly resolving both spectral and spatial (rather than temporal) information:

> the 2D receptive-field profiles of simple cells in mammalian visual cortex are well described by members of this optimal 2D [Gabor] filter family, and thus such visual neurons could be said to optimize the general uncertainty relations for joint 2D-spatial-2D-spectral information resolution.

So there is a distinctively mathematical, non-causal explanation of why the Gabor function is optimal, which can form part of a bigger causal explanation of why V1 neurons should have evolved (or developed) their observed properties.

Other neuroscientists have focussed on the causal factors leading to the development of Gabor-like RF's, and such enquiry does yield experimental interventions. For example, Hyvärinen and Hoyer (2001, 2413) suggest that, "[t]he reason why the CRFs [classical receptive fields] have Gabor-like shapes might thus be that these kind of CRFs are optimal for analyzing the input that the visual system typically receives". In other words, the hypothesis is that in evolutionary or developmental time, V1 RF's have been adjusted to be most sensitive to the kind of visual information usually prevalent in the natural environment. As it happens, short edges and bar like structures are very common in images of natural scenes, and so it seems plausible that the Gabor-like receptive fields are a good way to recover these kinds of stimuli. In order to test the hypothesis that in the course of development, neurons in V1 adjust their RF shapes to be most tuned to the common kinds of structures in the environment it is possible to do causal experiments in which animals are reared in particular environments (e.g. one which is full of vertical bars) and see if neurons with a preference for vertical stimuli then dominate V1. Such experiments were quite popular in the 1960's and 70's, and significant developmental effects could be shown [e.g. Blakemore and Cooper (1970)].

It should be noted that causal and non-causal explanations can be mutually supportive. One may demonstrate that a neural system bears a similarity to a theoretically optimal system but this could just be a coincidence. If by changing the circumstances in which the system lives, one changes what counts as the optimum for the system, and the system diverges from the previously observed behaviour in a predictable

way, that is good evidence that the previous behaviour was functionally significant and did not just bear a coincidental resemblance to the theoretical optimum. So any alternation between causal and non-causal explanation in neuroscience should not be taken as a troubling sign of theoretical double-think, but as a scientifically responsible strategy which profits from the insights of mathematical theorising, while keeping theoretical speculation grounded in empirical facts.

4 CASE III: A DYNAMICAL MODEL OF PREFRONTAL CORTEX

When addressing the challenge of understanding actual behaviours, rather than just single neurons, or general computational features of the brain, neuroscientists must reckon with the daunting complexity of neural systems. Multi-electrode neurophysiology allows researchers to listen in on the activity of 100's of neurons at a time, but even such small samples of activity from one brain area are extremely difficult to interpret functionally. Techniques of dimensionality reduction and dynamical systems analysis, imported from other branches of science, have become popular in the quest to simplify the brain. One recent study from Bill Newsome and Krishna Shenoy's labs at Stanford employs these techniques in order to shed light on the question of how the brain makes behavioural decisions based on complex sensory stimuli. It is worth considering whether their explanation of context-dependent decision making is causal, non-causal, or both. In order to address this question, it is first necessary to describe the study in a fair amount of detail.

4.1 A New Explanation of Context-Dependent Computation

The focus of enquiry was the flexibility of humans and other primates in responding to relevant sensory stimuli, depending on contextual cues. The behavioural task targeted this ability by presenting monkeys with a visual display of hundreds of red and green moving dots. The proportion of different coloured dots would change from trial to trial, as would the predominant direction of motion. On some trials, a cue indicated to the monkeys that they should give their response based on the majority colour, and on the other trials they were cued to respond based on the predominant direction of motion. Thus the same stimulus required different kinds of responses depending on a contextual cue.

During this task, recordings were taken from populations of neurons in prefrontal cortex (PFC), an area believed to be involved in controlling flexible, context-dependent behaviour. A number of models have been proposed to account for the neural basis of this behaviour, and one of the aims of the study was to test those models against the behavioural and neural data. The crucial step in the assessment of the models was the use of principle components analysis (PCA) as a means to simplify the neuronal population dataset so that it could be represented in a three dimensional space. Mante et al. (2013a, 79) interpret their three axes as corresponding to *choice* (which out of two responses the monkey will give), *motion* (the representation of information about the predominant direction of movement of the dots), and *colour* (the representation of information about the

predominant colour of the dots). The dynamical response of the population to the stimulus on any one trial is plotted as a state space trajectory in these three dimensions. One key finding was that whether or not the contextual cue is for a colour or motion response, the irrelevant sensory information is still represented by the neuronal population. This rules out one popular, *early selection* model of context-dependent computation, which assumes that the irrelevant information is filtered out before it reaches PFC. Another key finding was that the orientation of the axes relative to the others remains fixed, regardless of which context is used. This rules out two other models, which assume that the choice axis lines up with the axis of relevant sensory information, leading to the contextually appropriate response (Mante et al., 2013a, 81).

Given the failure of these existing models to account for the data, the authors go on to present their own model, which they describe as, "a previously unknown mechanism for selection and integration of task-relevant inputs" (Mante et al., 2013a, 78). The first step is the training of an artificial network of "recurrently connected, nonlinear neurons" (Mante et al., 2013a, 81) to perform a simulated version of the behavioural experiments, and see if the artificial "data" have the same qualitative features of the real neuronal data. They then analyse the dynamics of the trained, artificial network to see if they can account for these features.

In the three dimensional state space representation, the model population did indeed have the same qualitative features as the neuronal population: irrelevant sensory information was still represented at the population level, and the position of the choice axis remained fixed relative to the other axes (Mante et al., 2013a, 82). The use of the artificial network allowed them to perform a simulated experiment not possible with the real neuronal population, which is to perturb the network into a particular state, and see what state the network relaxes into. This reveals the fixed points of the network dynamics—the states which the network always relaxes back to. The analysis showed a series of approximate fixed points—known as a *line attractor*—along the choice axis. Another feature of this dynamical system, referred to as the *selection vector*, was that perturbation of the network in the direction of the *relevant* stimulus dimension caused the network to relax to a point along the line attractor which is closer to a final "choice point", whereas perturbation of the network in the direction of the *irrelevant* stimulus dimension resulted in the network just relaxing back to its initial point on the line attractor. The authors interpret this feature of the system as a "mechanism" for the integration of relevant sensory information (Mante et al., 2013a, 83).

4.2 Causal or Non-causal?

It is interesting that the authors of the study repeatedly refer to these features of the network dynamics—the line attractor and the selection vector—as indicating a possible mechanism for context dependent computation in PFC. The important point is, however, that the "components" of the posited mechanism are features revealed only by the dynamical analysis.[26] This

26 Note that because the fixed points, etc., are only revealed by reverse engineering the model through perturbation and relaxation, rather than through an equivalent experiment on the real neural network (which would not be possible with current techniques), the authors are careful to present it as a possible or plausible explanation of the behaviour of PFC, rather than a "how-actually" model of this brain area. However, it can be understood as an actual explanation of the artificial neural network, so for the purpose of this section I will just focus on

means that the notion of mechanism in play here is rather different from the one usually employed by philosophers of neuroscience and it is not at all a foregone conclusion that the kind of explanation offered by this model is a causal-mechanistic one.

In their discussion of the use of dynamical models to give causal-mechanistic explanations, Kaplan and Bechtel (2011, 439 and 443) claim that this is possible if the equations of the model can be interpreted as describing the activities of parts of actual neurons or circuits. This is consistent with the "models to mechanism mapping" (3M) criterion presented by Kaplan (2011, 347) and Kaplan and Craver (2011, 611). However, the fixed points which Mante et al. (2013a) characterise as explaining the context dependent computations cannot be thought of as directly mapping onto any actual parts of a real or artificial neural network; rather, they are attractors in an abstract, low dimensional state-space.

It is worth comparing the PFC model with a dynamical model more obviously amenable to the mechanistic account. Bechtel (2011, 553) presents the example of the dynamical model of circadian rhythms in *Drosophila*, writing that, "the equations are advanced ... as descriptions of the operations of specific parts of a mechanism" and that, "an important part of evaluating the adequacy of a computational model is that the parts and operations it describes are those that can be discovered through traditional techniques for decomposing mechanisms". In this case it is straightforward to advance a model-to-mechanism mapping: the equations in the model represent such things as the rate of change in concentration of mRNA of a particular gene (*per*) in terms of a rate of transcription and rate of decay. A computational simulation then reveals the dynamics of the system described by the equations, and the fact that both simulation and biological system manifest the same oscillatory pattern is reason for thinking that the dynamical model of the components and operations of the cell (the set of differential equations) explains circadian rhythms. Here, the issue of how to interpret highly abstract mathematical objects, such as line attractors, does not arise.

In order to advance a conventional mechanistic interpretation of the PFC model, one would have to read "organizational features of the target mechanism" very liberally, such that the term refers to some non-localisable and non-decomposable features of the network, ones which can only be revealed by DST analysis, rather than components and activities which are observable without such techniques. Only in this way could the 3M criterion be satisfied. Yet this move invites the objection that this very liberal application of 3M stretches the notion of mechanistic explanation beyond recognition, and indeed usefulness.

On the view of Silberstein and Chemero (2013, 967), "global organizational principles or features of complex systems" which are revealed by DST analysis, "are not explicable in principle via localization and decomposition", and thus should not be interpreted mechanistically. The failure of the mechanistic heuristics of localisation and decomposition occurs, they argue, because many structurally diverse networks which differ in their implementational details still exhibit the same global features. Note that the same independence from implementional details is true for the artificial neural networks described in the PFC study—the modellers

the explanation of the behaviour of the model itself. As Sussillo and Barak (2013, 627) note, the operation of these recurrent neural networks is often opaque to modellers because the training of the networks does not specify how the task should be performed. Thus it is necessary to open the black box using analytical techniques of DST.

trained 100 different networks, with different initial conditions and therefore different 'synaptic' connectivity patterns, and yet the global dynamical features were the same in each case (Mante et al., 2013b, 23).

So is the model explanatory in another, non-mechanistic, or even non-causal sense? One way to address this question is to see how closely the study resembles the kind of modelling practices used in physics and engineering (e.g. nonlinear fluid dynamics), which are often taken to be a exemplary of non-causal explanation. After all, many of the analytical tools employed in the PFC study were first developed in that context (Ott, 2002). Batterman (2002, 23) introduced the notion of "type-ii why questions" as those which require an explanation as to why a collection of micro-physically very different substances all exhibit the same macro-behaviour, e.g. during phase transitions. Such explanations are provided by mathematical abstraction techniques, such as the renormalisation group, which demonstrate that all of the different substances are members of the same "universality class", despite their low-level differences.

Ross (2015) shows that an analogous pattern of explanation occurs when dynamical models and abstraction techniques are used in cellular neuroscience, to explain why anatomically very different neurons all exhibit the same spiking behaviour, described by a "canonical model". Now one could argue that a similar explanation is provided by the PFC study because the authors demonstrated that 100 different networks trained to perform the context-dependent decision, all with different patterns of synaptic connections, converged on the same high level dynamical properties (line attractor and selection vector), which were revealed by the abstraction techniques of PCA and the perturbation-relaxation investigation.

However, one must also consider that there is a *functional* explanation for the different networks' convergence on the same dynamics. Mante et al. (2013b, 23) write that,

> We trained many networks (around 100) from different initializations of the weights and biases, and each time the network solved the problem in the same qualitative way. This points to the fact that the selective integration task . . . placed strong constraints on the optimization process.

The implication is that the demands of the task causally shaped the development of the network in such a way, through the optimization process, that all the different networks inevitably converged on a pattern of connections which could implement the high-level dynamics and thus solve the context-dependency problem. It is an interesting question whether the non-causal explanation of universality, or the causal-functional explanation of convergence onto the same behaviour is deeper, or more scientifically significant in this case; or indeed whether the causal and non-causal explanations complement each other, as I argued they do in the case of the Gabor model.

Lacking firm grounds to answer this question, or even clear intuitions either way, I will bracket the issue for now. As I see it, the more important explanandum which the PFC study seeks to address is *how it is that any one network can perform the context-dependency task*—rather than the type-ii why question over why different networks perform the task in the same way. Is the explanation here a non-causal one? Given that we still have open questions about the interpretation of dynamical models, one thing we can safely say is that the answer to this question turns on how we

choose to interpret the working parts of the explanation (i.e. the line attractor and selection vector). If one regards these as somehow capturing or summarising the actual causally-efficacious parts of the network (the properties of real and artificial neurons), then one would be inclined to say that the explanation is a causal one. For instance, one could say that causal properties of the network which allow it to integrate information are modelled and represented by the mathematical artifice of the line attractor, where the notion of mathematical representation without explanation is invoked (Saatsi, 2011). On the other hand, if one finds such an interpretation far fetched, then it is natural to say that the explanation is a non-causal one.

This brings us to the point that arguably the PFC model is not explanatory at all. The neuroscientists' account of context dependent computation is entirely reliant on the three principle axes revealed by PCA, and the interpretation of these as indicating the "choice" of the network, alongside motion and colour information. This raises the thorny question of how we should interpret the results of factor analysis. Some might accuse Mante and colleagues of committing the sin of reifying their essentially meaningless primary factors—that is, of "awarding *physical meaning* to all strong principal components" (Gould, 1981, 250).

In sum, we have three options on the table: the first is to say that the model offers a causal-mechanistic explanation; the second that it provides some sort of non-causal explanation; the third position is that it is not an explanatory model. What reason is there to favour the second position? We can begin by highlighting problems with the first and third options.

I am not moved by the view that the PFC model is not at all explanatory; this attitude strikes me as being tone deaf to the scientifically compelling features of the analysis. Because of the way that their network moves through a low dimensional space in response to "sensory" input, the researchers are able to show that it is able to integrate relevant information and ignore irrelevant information, while retaining information about both kinds of stimuli, as observed in the actual neural data. The explanandum is, 'how is this brain area able to integrate relevant sensory information?', and by being shown specific features of the global dynamics we are indeed provided with an explanans. If one is going to be skeptical about the meaningfulness of principal components in this case, one ought to be unpersuaded by any explanations which call upon statistical constructs, such as the greater profitability of Sunnylake Farm, compared to the otherwise identical Sunnybrook Farm, being due to the higher mean number of eggs laid per day.[27]

One strike against the mechanistic interpretation is that the explanation offered by the dynamical model of PFC is *not* a constitutive one. That is to say, it is *not* an explanation of how a global phenomenon—computation—comes about because of the activities of some network components, whether spatially localised or not. Instead, the dynamical properties of the network, which do the job of explaining the computational phenomena, are themselves global or population level properties of the network.[28]

27 As Gould (1981, 250) notes, the warning against reification of principal components is not intended as a blanket ban. Sometimes interpretation is justified by our knowledge of how the system behaves and what we are likely to be measuring.

28 See Mante et al. (2013a, 79): "To study how the PFC population as a whole dynamically encodes the task variables underlying the monkeys' behaviour, we represent population responses as trajectories in neural state space. Each point in state space corresponds to a unique pattern of neural activations across the population." Also Mante et al. (2013a, 78): "The mechanism reflects just two learned features of a dynamical system: an approximate line attractor and a 'selection vector', *which are only defined at the level of the population*" (emphasis added). So my

Another important point is that the model gives us no information about how we might make changes to the network in order to affect changes to its information processing properties. It does not tell us which connections would have to be rearranged in order to make the computation no longer context dependent, for example. However, one might ask, doesn't the model tell us how we could intervene on the *global* dynamical properties of the network (rather than local connectivity patterns) in order to bring about changes in its computational properties? Isn't it therefore a pared down representation of some causal relationships?[29] I believe that we should not interpret the counterfactual dependency between global dynamical features and computational properties as a causal relationship, even in a very minimal sense. This is because the dynamical description, and the descriptions of the network's computational properties are just two different ways of describing the same thing—the network. There is no spatio-temporal separation between putative cause and putative effect. To change the dynamical properties is just to change the information processing capacities of the network.

This brings me to say more about the second option which I endorse, i.e. that the model offers some kind of non-causal explanation. We can think of the dynamical model as offering an illuminating *perspective* on the network. It makes transparent certain counterfactual relationships holding between the network's global dynamical features and the computations which it performs—for instance, that the ability of the system to integrate sensory information for the duration of a trial is couterfactually dependent on it having a line attractor rather than a point attractor. Thus the model answers w-questions. We can say, for example, that if the selection vector were never orthogonal to the irrelevant stimulus dimension, then that information would not be ignored. This is another instance of an explanatory model providing counterfactual information which should not be interpreted as describing the outcomes of possible interventions.

So if my interpretation is correct, this example fits into Woodward's schema for non-causal explanation. Is the explanation also distinctively mathematical in Lange's sense? Well, if it can be shown that there is a modally strong connection between the having of certain dynamical properties (one kind of mathematical description) and the ability to perform certain computations (another kind of mathematical description), then this example would also satisfy Lange's criteria. But I leave this matter for readers with greater technical expertise than I myself have. One final point is that thinking of the model as providing explanations by offering illuminating global perspectives on a system helps shed light on why Batterman and Rice's "minimal model explanations" are correctly described as non-causal ones. The point is not just, as they suggest, that minimal models like the DST one only offer caricatures of the causal dependencies [Batterman and Rice (2014), cf. Rice (2015)]. More strongly, when the features captured by the model are truly global, population-level ones, then no causal dependencies are being represented at all—either in an accurate or caricatured way. Instead, what the model does is reveal new, global, features of the system which indicate why the system also has the global, observable features which have served as our explananda.

point is that what these authors call a "mechanism" for context dependent computation cannot be thought of as providing a constitutive explanation.

29 I.e. a very roughly sketched description of difference makers [Weisberg (2007), Strevens (2008)].

5 CAUSAL AND NON-CAUSAL: DOES THE DIFFERENCE MATTER?

In this paper I have shown that there are instances in computational neuroscience of explanations which are precisely analogous to familiar examples of non-causal explanations in physics and biology. So if one is persuaded by Lange's account of non-causal, distinctively mathematical explanation, or by the non-causal extension of Woodward's counterfactualist framework, then one should agree that computational neuroscience is just another domain in which such patterns of explanatory reasoning occur. What is interesting about this result is that it stands against the dominant mechanistic current of recent work in the philosophy of neuroscience, which seeks to interpret explanatory practice as the mapping of mechanisms: working out what cause produces which effect, in order to show how neuronal components are orchestrated to govern complex behaviours, such as navigation. Moreover, by extending Woodward's framework to incorporate distinctively mathematical explanation I am adapting materials frequently used by mechanists to the rather different task of analysing non-causal explanation.

The clearest cases of non-causal explanation in neuroscience are *efficient coding explanations* which refer to information theoretic trade-offs in order to show why it is that neural systems should employ particular computational solutions, such as hybrid computation (Section 2), or Gabor filtering (Section 3). So far most of the discussion of non-mechanistic explanation in neuroscience has focussed on dynamical systems theory. Yet as I show in Section 4.2, the argument that explanatory practice goes beyond the detailing of mechanisms is less clear cut in this area. The controversy turns on the question of whether or not we should interpret the mathematical structures revealed by dynamical modelling (e.g. fixed points and line attractors) as representing anything in the brain which has causal powers to affect computations performed.

So does the difference between causal and non-causal explanation matter to neuroscientists and neuroengineers themselves? Perhaps not. As noted at the end of the discussion of the Gabor model, these different kinds of explanation should be seen as complementary rather than in competition with one another [Cf. Andersen (forthcoming)].

Still, one might point out that if what we are interested in explaining is why some empirically realized system (like the brain) exhibits some feature, then this explanation must always have an empirical component (e.g. that some optimizing selective process is operative). So one objection to my account is that the purely mathematical part of the explanation is not really by itself an explanation of anything empirical and that one needs the component about natural selection (or something similar) in order to arrive at an explanation of the empirically observed facts.

In response I would agree that the non-causal component is arguably not a stand-alone explanation of the efficiency of neural computation or the Gabor shaped receptive fields of V1, even though researchers in theoretical neuroscience do talk of these mathematical results as explaining brain features. Such talk assumes that certain empirical background conditions obtain, such that the trade-offs are relevant to the neural systems, and that there is some developmental, evolutionary or realtime adaptational process by which the near-optimal solution is arrived at; but such a story is bracketed in order to focus on the relevant topic of enquiry, which is the

interpretation of a particular feature in terms of its utility and efficiency. So what is interesting here is that there is a division of explanatory labour: some neuroscientists will focus on the non-causal, mathematical explanation of the efficiency of a feature while it is the job of others to find out about the aetiology of that feature.

For the purposes of this paper I need not take a stand in the debate about the possibility of stand-alone non-causal explanation of empirical facts, or answer the question of whether we can call an explanation non-causal if there is some non-causal part of the story (or causal if there is some empirical part of the explanation). But it is worth pointing out that Lange's examples, such as the honeycomb one, invite the same objection. The task of this paper is to show that there are examples of explanation in neuroscience that are of the same type as the central examples of non-causal explanation as presented in the work of Lange, Batterman and Woodward. I hope by now to have made that case.

Even if I concede that my three cases do not give us stand-alone non-causal explanations, my examples are still relevant to the ongoing debate over whether there are distinctly computational and non-mechanistic patterns of explanation in neuroscience. While authors such as (Kaplan, 2011) echo Saatsi (2011) and argue that mathematics only has the role in neuroscience of representing physical systems via a mapping from equations to parts and processes of a mechanism, I have argued that mathematics has a stronger explanatory role which is independent of mechanism description. We arrive at a picture in which theoretical neuroscience is often focussed on the non-causal parts of explanations and therefore has explanatory norms—appealing to efficient coding and mathematical trade-offs—that are distinct from the evolutionary and mechanistic branches of neuroscience. This, I believe, is sufficient for me to deflect the worry that efficient coding explanations are just another kind of evolutionary explanation, which are themselves elliptical mechanistic explanations.

One issue which I have not addressed is the question of how we evaluate the quality of non-causal explanations. One of the virtues of the mechanistic account is that it lays down explicit normative criteria (Craver, 2007). There is not space here to venture into this topic, but one obvious possibility is to evaluate non-causal explanations according to the precision and range of counterfactual or counterpossible information provided (i.e. how precisely and extensively they are able to answer w-questions), in the same way that one can evaluate causal explanations.

A novel possibility is to consider that non-causal explanation tends to exemplify a different explanatory virtue from the causal sort, that of unifying diverse phenomena under one general principle (Kitcher, 1989). Though the unificatory account of explanation is less popular than it once was, it should not be ignored that all the cases I have presented do a good job of explaining why mechanistically quite diverse systems will converge on similar computational solutions—in other words, answering Batterman's type-ii why questions. In this era of big data neuroscience and relative scarcity of theoretical insight, unificatory explanatory knowledge is a precious commodity. It will be instructive to see whether future developments in neuroscience will place non-causal explanations in increasingly prominent roles.

REFERENCES

Andersen, H. (forthcoming). Complements, not competitors: causal and mathematical explanations. *British Journal for the Philosophy of Science*.

Attwell, D. and S. B. Laughlin (2001). An energy budget for signalling in the grey matter of the brain. *Journal of Cerebral Blood Flow and Metabolism 21*, 1133–1145.

Barberis, S. D. (2013). Functional analyses, mechanistic explanations, and explanatory tradeoffs. *Journal of Cognitive Science 14*, 229–251.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory Communication*. Cambridge, MA: MIT Press.

Baron, S., M. Colyvan, and D. Ripley (forthcoming). How mathematics can make a difference. *Philosopher's Imprint*.

Batterman, R. (2002). *The Devil in the Details*. Oxford: Oxford University Press.

Batterman, R. (2010). On the explanatory role of mathematics in empirical science. *British Journal for the Philosophy of Science 61*, 1–25.

Batterman, R. and C. Rice (2014). Minimal model explanations. *Philosophy of Science 81*(3), 349–376.

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

Bechtel, W. (2011). Mechanism and biological explanation. *Philosophy of Science 78*(4), 533–557.

Bjerring, J. C. (2014). On counterpossibles. *Philos Stud 168*, 327–353.

Blakemore, C. and G. F. Cooper (1970). Development of the brain depends on the visual environment. *Nature 228*, 477–478.

Bokulich, A. (2008). Can classical structures explain quantum phenomena? *British Journal for the Philosophy of Science 59*, 217–35.

Bokulich, A. (2011). How scientific models can explain. *Synthese 180*, 33–45.

Brogaard, B. and J. Salerno (2013). Remarks on counterpossibles. *Synthese 190*, 639–660.

Chemero, A. and M. Silberstein (2008). After the philosophy of mind: Replacing scholasticism with science. *Philosophy of Science 75*, 1–27.

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

Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory* (2nd ed.). Hoboken, NJ: Wiley Interscience.

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

Craver, C. F. (2014). The explanatory power of network models. *PSA presentation*.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am. A 2*(7), 1160–1169.

Gabor, D. (1946). Theory of communication. *Journal of the Institution of Electrical Engineers 93*, 429–459.

Godfrey-Smith, P. (2001). *Three Kinds of Adaptationism*, pp. 335–357. Cambridge: Cambridge University Press.

Gould, S. J. (1981). *The Mismeasure of Man*. London: Penguin.

Huneman, P. (2010). Topological explanations and robustness in biological sciences. *Synthese 177*, 213–245.

Husbands, P. and O. Holland (2008). The ratio club: A hub of british cybernetics. In P. Husbands, O. Holland, and M. Wheeler (Eds.), *The Mechanical Mind in History*, pp. 91–148. MIT Press.

Hyvärinen, A. and P. O. Hoyer (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research 41*, 2413–2423.

Irvine, E. (2014). Models, robustness, and non-causal explanation: a foray into cognitive science and biology. *Synthese DOI 10.1007/s11229-014-0524-0*.

Izhikevich, E. M. (2010). *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. Cambridge, MA: MIT Press.

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

Kaplan, D. M. and W. Bechtel (2011). Dynamical models: An alternative or complement to mechanistic explanations? *Topics in Cognitive Science 3*, 438–444.

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

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher and W. Salmon (Eds.), *Scientific Explanation*, pp. 410–505. Minneapolis: University of Minnesota Press.

Komornicki, A., G. Mullen-Schulz, and D. Landon (2009). *Roadrunner: Hardware and Software Overview*. http://www.redbooks.ibm.com/redpapers/pdfs/redp4477.pdf.

Lange, M. (2013). What makes a scientific explanation distinctively mathematical? *British Journal for the Philosophy of Science 64*, 485–511.

Levy, A. (2014). What was hodgkin and huxley's achievement? *British Journal for Philosophy of Science 65*, 469–492.

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

MacKay, D. (1991). *Behind the Eye*. Oxford: Blackwells.

MacKay, D. and W. McCulloch (1952). The limiting information capacity of a neuronal link. *Bulletin of Mathematical Biophysics 14*, 127–135.

Mante, V., D. Sussillo, K. V. Shenoy, and W. T. Newsome (2013a). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature 503*, 78–84.

Mante, V., D. Sussillo, K. V. Shenoy, and W. T. Newsome (2013b). Supplementary information.

McCulloch, W. and W. Pitts (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics 7*, 115 – 133.

Ott, E. (2002). *Chaos in dynamical systems*. Cambridge: Cambridge University Press.

Piccinini, G. and S. Bahar (2013). Neural computation and the computational theory of cognition. *Cognitive Science 34*, 453–488.

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

Rice, C. (2012). Optimality explanations: a plea for an alternative approach. *Biology and Philosophy 27*, 685–703.

Rice, C. (2015). Moving beyond causes: Optimality models and scientific explanation. *Noûs 49*(3), 589–615.

Rieke, F., D. Warland, R. d. R. V. Steveninck, and W. Bialek (1999). *Spikes: Exploring the Neural Code*. Cambridge, MA: MIT Press.

Ross, L. N. (2015). Dynamical models and explanation in neuroscience. *Philosophy of Science 82*(1), 32–54.

Saatsi, J. (2011). The enhanced indispensability argument: Representational versus explanatory role of mathematics in science. *British Journal for Philosophy of Science 62*, 143–154.

Saatsi, J. and M. Pexton (2013). Reassessing woodward's account of explanation: Regularities, counterfactuals, and noncausal explanations. *Philosophy of Science 80*(5), 613–624.

Sarpeshkar, R. (1998). Analog versus digital: Extrapolating from electronics to neurobiology. *Neural Computation 10*, 1601–1638.

Sarpeshkar, R. (2010). *Ultra Low Power Bioelectronics*. Cambridge: Cambridge University Press.

Serban, M. (2015). The scope and limits of a mechanistic view of computational explanation. *Synthese 192*(10), 3371–3396.

Silberstein, M. and A. Chemero (2013). Constraints on localization and decomposition as explanatory strategies in the biological sciences. *Philosophy of Science 80*(5), 958–970.

Stepp, N. A., Chemero, and M. T. Turvey (2011). Philosophy for the rest of cognitive science. *Topics in Cognitive Science 3*(2), 425–437.

Sterling, P. and S. B. Laughlin (2015). *Principles of Neural Design*. Cambridge, MA: MIT Press.

Strevens, M. (2008). *Depth: An account of scientific explanation.* Cambridge, MA: Harvard University Press.

Sussillo, D. and O. Barak (2013). Opening the black box: Low-dimensional dynamics in high-dimenional recurrent neural networks. *Neural Computation 25*, 626–649.

von Neumann, J. (2000). *The Computer and the Brain*. New Haven: Yale University Press.

Weisberg, M. (2007). Three kinds of idealization. *Journal of Philosophy 104*(12), 639–659.

Weiskopf, D. A. (2011). Models and mechanisms in psychological explanation. *Synthese 183*, 313–338.

Woodward, J. F. (2003). *Making Things Happen*. New York: Oxford University Press.

Woodward, J. F. (2013). Mechanistic eplanation: its scope and limits. *Proceedings of the Aristotelian Society Supplementary Volume 87*, 39–65.

Woodward, J. F. (2014). A functional account of causation. *Philosophy of Science 81*(5), 691–713.

Woodward, J. F. (forthcoming). Explanation in neurobiology: An interventionist perspective. In D. M. Kaplan (Ed.), *Integrating Psychology and Neuroscience: Prospects and Problems*. Oxford: Oxford University Press.

Wouters, A. G. (2007). Design explanation: determining the constraints on what can be alive. *Erkenntnis 67*, 65–80.