THE DEVELOPMENT AND APPLICATION OF EFFICIENT CODING EXPLANATION IN NEUROSCIENCE

M. CHIRIMUUTA*

August 17, 2016

CONTENTS

1 Introduction 2
2 Efficient Coding Explanation and the Causal/Non-Causal Frontier 4
3 Lateral Inhibition and Explanations of Early Visual Responses 8
3.1 Edge Detection and Feature Sharpening 10
3.2 Redundancy Reduction 12
3.3 Predictive Coding 16
3.4 Some Observations 18
4 Putting Efficient Coding Explanations to Use 18
4.1 Scaling Data Mountain 18
4.2 Forward Engineering 20
4.3 Defining Neural Computation 21
5 Conclusion 22

ABSTRACT

Theories of causal explanation are often championed because they make obvious the connection between scientists' explanatory practices and their ability to manipulate objects and processes in the natural world. In the philosophy of neuroscience, much attention has been paid to mechanistic explanations, both in terms of their theoretical virtues, and their application in potential therapeutic interventions. Non-mechanistic, non-causal

* mac289@pitt.edu.

This is the penultimate version of a chapter forthcoming in *Explanation Beyond Causation* (eds. Alex Reutlinger and Juha Saatsi, Oxford University Press). I would very much like to thank Peter Sterling and the editors of the volume for many thoughtful comments which have helped me to improve this contribution.

Note to reader: in order to meet the word limit the final version of the chapter leaves out much of the detail on the development of theories lateral inhibition, and most of the long quotations are gone. I've left them in this version, in case you are interested in following up the historical material. This version was not thoroughly edited, so awkward sentences and typo's abound.

explanatory models, it is often assumed, would have no role to play in any practical endeavours. This assumption ignores the fact that many of the non-mechanistic, explanatory models which have been successfully employed in neuroscience have their origins in engineering and applied sciences, and are central to many new neuro-technologies, such as brain computer interfaces. In this chapter I present a case study of the development of explanations of *lateral inhibition* in the early visual system as implementing an *efficient code* for converting photoreceptor input into a data-compressed output from the eye to the brain. I discuss two applications of the efficient coding approach: in streamlining the vast datasets of current neuroscience by offering unifying principles, and in building artificial systems that replicate vision and other cognitive functions. I also argue that efficient coding models can fruitfully be employed in the task of defining neural computation.

1 INTRODUCTION

Recent philosophy of neuroscience (since circa 2000) has been dominated by discussion of mechanisms. The central proposal of work in this tradition is that explanations of the brain are crafted through the discovery and representation of mechanisms. Another core commitment is to explanation being a matter of situating phenomena in the causal structure of the world. This is often accompanied by a commitment to an interventionist theory of causation and causal explanation. Accordingly, a criterion of explanatory sufficiency is the ability of a theory or model to tell us how our phenomenon would be altered under different counterfactual scenarios—the ability to answer *what-if-things-had-been-different-* or *w*-questions (Woodward, 2003).

Various authors believe that 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. One thing that might seem puzzling about this move is that it extends Woodward's framework in such a way as to apparently divorce scientific explanation from the demands of working out how to intervene successfully in the world. The tight connection between causally explaining and making a difference was originally one of the selling points of Woodward's account (Reutlinger, 2012). Yet if an explanation fulfills the counterfactualist, but not the interventionist norms, it can seem hard to find a point to the investigation beyond theoretical speculation. For when one learns of a non-causal explanation of, say, patterns of spiking and non-spiking activity in a neuron, one is not thereby learning of the specific "levers and pulleys" which would allow one to impede a pathological kind of neuronal behaviour, such as underlies epileptic disease.

See Bokulich (2008), Bokulich (2011), Saatsi and Pexton (2013). As Woodward (forthcoming, 5) puts it, "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". If the "changes" are brought about by interventions, then we have causal explanation; if they cannot be understood in this way (e.g. because they involve changes in the laws of mathematics) then we have non-causal counterfactualist explanation.

I thank Anna Alexandrova for raising this issue. Even though the interventionist theory of causation only need refer to hypothetical interventions, not actual ones, advocates of interventionism often highlight the connection between this way of thinking about causation and the practice of figuring out ways to alter the course of natural events. E.g. Kaplan and Craver (2011, 602).

I have recently argued that the w-question criterion can be satisfied by models of neural systems which are non-mechanistic (Chirimuuta, 2014) and non-causal (Chirimuuta, forthcoming). I refer to these as *efficient coding explanations*. Such explanations occur frequently in computational neuroscience—a broad research area which uses applied mathematics and computer science to model neural systems. The models in question ignore biophysical specifics in order to describe the information processing capacity of a neuron or neuronal population. 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* (See Section 3.2 and Footnote). 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.

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 running the system. 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 Attwell and Laughlin (2001, 1133) write:

> The neural processing of information is metabolically expensive. Although the human brain is 2% of the body's weight, it accounts for 20% of its resting metabolism .... This requirement for metabolic energy has important implications for the brain's evolution and function. The availability of energy could limit brain size, particularly in primates ..., and could determine a brain's circuitry and activity patterns by favoring metabolically efficient wiring patterns.

It is important to note that efficient coding explanations do not rely on the strong adaptationist assumption that the brain of humans, or any other animal, *is* somehow optimal. 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. Barlow (1961a, 224) gives a useful statement of this methodological policy:

> the safe course here is to assume that the nervous system is efficient. If it is clearly demonstrated that the nervous system

This passage from Doi et al. (2012, 16256) encapsulates the idea: "It has been hypothesized that the early stages of sensory processing have evolved to accurately encode environmental signals with the minimal consumption of biological resources .... This theoretical hypothesis, generally known as efficient coding, has been used to explain a variety of observed properties of sensory systems." And see references therein.

> is inefficient in some particular well-defined way, this can quite easily be incorporated into the hypothesis and its implications correspondingly modified, whereas our whole frame of thought might be undermined if it turned out that the nervous system was more efficient than we had supposed.

In this chapter I argue that efficient coding explanations have important roles to play in various kinds of practical activity. There are more ways to make a difference than facilitating and preventing causal effects; one may also wish to build things. There is a close and historically embedded connection between engineering and the research traditions in neuroscience which typically employ efficient coding reasoning. Thus we find numerous instances of efficient coding reasoning in attempts both to reverse engineer the nervous system and to forward engineer devices which replicate some of the functions of the biological brain.

In the next section I will outline some of the specifics of efficient coding explanation, and present my criteria for non-mechanistic and non-causal explanation. After that I will focus on the specific case of models of *lateral inhibition* in the early visual system (Section 3), following with discussion of two important applications: in scaling the so-called data mountain (Section 4.1) and in building artificial computational systems (Section 4.2). I also argue in Section 4.3 that models of this sort can fruitfully be employed in the task of defining neural computation.

2 EFFICIENT CODING EXPLANATION AND THE CAUSAL/NON-CAUSAL FRONTIER

Holly Andersen offers many useful reflections on the much contested frontier between causal and non-causal explanation. Causal explanation is often defined broadly as the placing of the explanandum phenomenon within the network of causal relationships in the world. A more stringent definition asserts that for an explanation to be causal, the connection between the explanans and explanandum must be a causal one (Andersen, forthcoming, 4). This would rule out constitutive mechanistic explanation, since in those cases the relationship between the entities and activites of the explanans, and the explanandum phenomenon, is one of constitution rather than causation. This strikes me as a problematic feature of Andersen's narrow definition. As I see it, constitutive mechanistic explanations where both explanans and explanandum are characterised as a set of causal relationships, should count as a kind of causal explanation. The important point is that the explanandum is *doing* something which brings about the explanans. As Kaplan and Craver (2011, 611) put it, it is important to see how the mechanism "produces, maintains, or underlies the phenomenon".

The lesson here is that there *is* a difference worth marking between mechanistic and aetiological explanation, but that does not mean that mechanistic explanation is non-causal—it is simply a different kind of causal explanation. By focussing on the relationship between explanans

For more on the historical links, see Husbands and Holland (2008) on the Ratio Club (1949-1958). This was a small scientific society consisting of neurobiologists, mathematicians, psychiatrists and computer engineers who were concerned with application of the new formalisms and concepts of information theory and cybernetics in the understanding of brain and behaviour. Donald MacKay and Horace Barlow were two members who have had a foundational influence on theoretical and computational neuroscience.

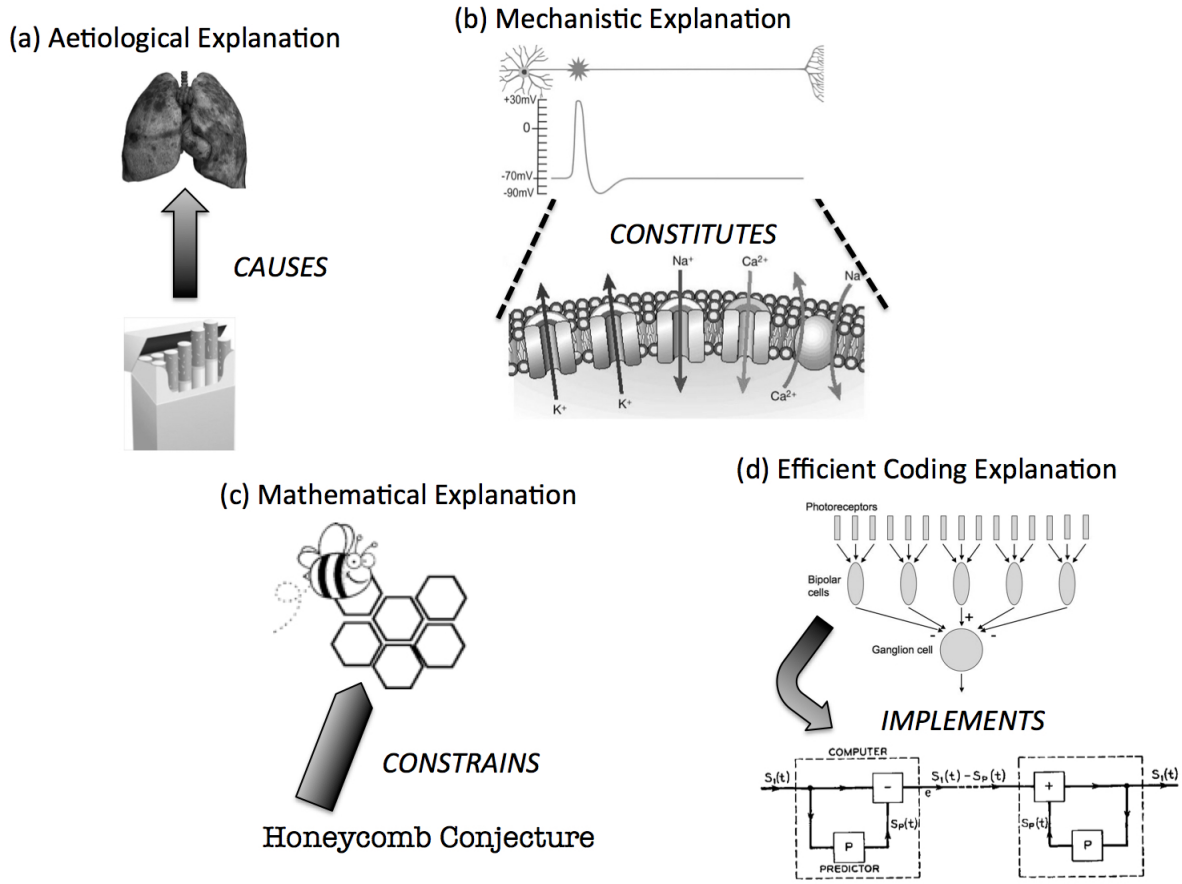

Figure 1: Four Kinds of Explanation. In each case the explanandum, depicted at the top, is a biological phenomenon. (a) Aetiological Explanation: smoking is said to *cause* the explanandum (lung disease). (b) Mechanistic Explanation: ion channels opening and closing in response to changing membrane potential is said to *constitute* the explanandum (action potential). (c) Mathematical Explanation: the explanans is a mathematical fact (the Honeycomb Conjecture) and it can be thought of as *constraining* the path of evolution towards the optimal solution to the bees' storage problem. (d) Efficient Coding Explanation: the explanans is an abstract coding scheme which is said to be *implemented* by the actual retinal circuit.

and explanandum we can chart these and other kinds. In each of the four examples depicted in Figure 1, the explanandum is a biological phenomenon. In the cases of (a) aetiological and (b) mechanistic explanation, the explanantia are also phenomena which can naturally be described as a series of causal processes. I classify these as two species of causal explanation. In (c) we have the non-causal, mathematical explanation of the hexagonal shape of honeycomb. The explanans is a law or fact of mathematics—the honeycomb conjecture proved by Hales (2001)—rather than an empirically observable causal process. One cannot speak of the mathematical facts as causing anything to happen in nature, though they do *constrain* the sequence of biological events.

The fourth kind, (d) efficient coding explanation, is clearly different from mechanistic and aetiological explanation in that the explanans is an abstract coding scheme or algorithm, rather than an empirically observable causal process. Also, the relationship between explanans and explanandum is one

of *implementation* rather than causation or constitution. Thus I classify it as a kind of non-aetiological and non-mechanistic explanation. One important point, though, is that unlike in the case of mathematical explanation where the explanans must be a "modally strong mathematical fact", it is often natural to think of the coding scheme in quasi-causal terms because it describes an input-output function and the series of steps needed to go from on to the other. That said, the coding scheme or algorithm is not a set of empirically observable causal relationships, and can also be thought of as an abstract mathematical object. Thus I am reluctant to classify computational explanation, as a general kind, as distinctively mathematical or either causal or non-causal.

In the cases of efficient coding explanation that I will discuss in this chapter, the neural system is said to implement a specific code or coding strategy, and this reasoning yields insights into why the system behaves in the ways observed. To take an example which I discuss at greater length elsewhere (Chirimuuta, forthcoming, §2), it has been argued that the nervous system implements *hybrid* computation—a manner of processing information which alternates between analogue and digital codes (Sarpeshkar, 1998). One property of hybrid computation is that it is energy efficient, using little power for each bit of information processed, in comparison with digital computation, while being less easily impacted by noise than purely analogue computation. Sarpeshkar argues that the implementation of hybrid computation explains how biological brains can consume orders of magnitude less energy than man-made supercomputers, while being equivalent in computational capacity.

Here the explanandum is a particular behaviour or feature of a neural system, namely the economy with which nervous tissue consumes energy. The explanans is a coding scheme, an abstractly characterised method of performing computations which has certain properties of its own, such as economical consumption of resources. There are mathematical frameworks, such as information theory, which tell us why the explanans has the property of interest. Physiological data are offered to provide evidence that the neural system *implements* the coding scheme. It is then argued that the reason why the neural system has the property of interest is that it is an implementation of the coding scheme theoretically shown to have this property. We then have an explanation of why the nervous tissue has the property in question.

This explanation is non-mechanistic because it does not proceed by decomposing the neural system and describing how the different component parts interact to give rise to the explanandum phenomenon [Machamer et al. (2000), Bechtel and Richardson (2010)]. This idea that mechanistic explanations work by tracing the causal relationships between components of a tightly knit biological system is also encapsulated in the "models to mechanism mapping" (3M) criterion:

> In successful explanatory models in cognitive and systems neuroscience (a) the variables in the model correspond to components, activities, properties, and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the perhaps mathematical dependencies posited among these variables in the model correspond to the

For the purposes of this paper I will bracket the vexed philosophical debate over the proper analysis of this term, noting that it is the concept of implementation is employed widely within neuroscience. But see Sprevak (2012) for an excellent discussion of the philosophical issues.

> perhaps quantifiable causal relations among the components of the target mechanism. [Kaplan and Craver (2011, 611), cf. Kaplan (2011, 347)]

The 3M criterion was introduced as part of an argument that all genuinely explanatory models in computational neuroscience are mechanistic ones. It is important to study efficient coding models because we find cases of explanation without 3M-style mapping (Chirimuuta, 2014, 145). For example, with hybrid computation, we are not told how particular components of the coding scheme relate to a neural system, as unearthed through physiological and anatomical study.

One might object that implementation is itself a kind of mapping relationship, and so efficient coding explanations satisfy the 3M criterion for mechanistic explanation. However, this argument misses the point that the central feature of mechanistic explanation is the tracing of causal relationships between the components of the explanans—the presentation of a mechanistic description—and showing how this set of relationships is responsible for some of the causal properties of the explanandum phenomenon. In the case of efficient coding explanation, the explanans itself (not just the representation of it) is a mathematical object, namely, a coding scheme or algorithm; the explanans is not a set of entities and activities in a biological system. Moreover, the relationship of implementation is not the constitutive one that is required for mechanistic explanation.We cannot say that the coding scheme "produces, maintains, or underlies" the neural phenomenon; instead, the neural system is just an instance of the coding scheme, realized in biological hardware.

Even if efficient coding explanations are non-mechanistic, one may still wonder if they are causal. Here things become a little complex. As has been noted elsewhere, when scientists present explanations of evolved systems which are subject to biological, physical, and mathematical laws, different kinds of explanations often rub-shoulders and one can shift between causal and non-causal explanations with subtle changes in the specification of the explanandum [Chirimuuta (forthcoming); Andersen (forthcoming)]. For example, the explanation of *why honeycomb is hexagonally shaped* must cite both the causal biological facts that there is evolutionary pressure on honeybees to maximize storage volume and minimise building materials in making comb, as well as the mathematical argument that a hexagonal structure is the one which achieves this aim. However, the explanation of *why honeycomb is the best structure, given the bees' needs* is "distinctively mathematical" (Lange, 2013, 499-500). In the case of hybrid computation, there is a causal (biological) explanation of why economy of computation is such an important factor in explaining nervous systems, whereas the explanation of why hybrid computation is optimal for biological brains is a non-causal one, based on principles of information theory (Chirimuuta, forthcoming, §2).

Before closing this section I would like to point out that all four kinds of explanation have the resources to answer what-if-things-had-been-different questions. In the case of mechanistic and aetiological explanation, we can conduct (real or hypothetical) experiments on the biological systems and observe how interventions on the explanans result in changes to the explanandum. While no-one could intervene on the laws of mathematics,

I say this because in the case of mechanistic explanation the mechanistic description may be presented as a mathematical equation, which is a representation of concrete entities and the causal processes occurring amongst them.

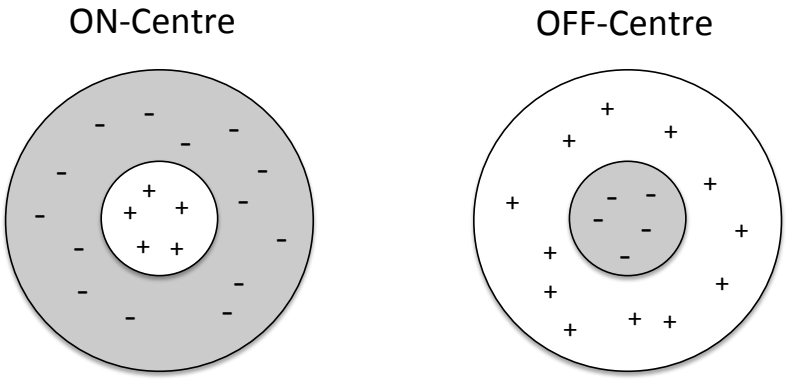

Figure 2: Receptive Fields of Retinal Ganglion Cells. If light falls on the excitatory centre of an ON cell, firing rate will increase, whereas rate decreases if light falls on the inhibitory surrounding area. The polarity of responses is reverse for OFF cells.

mathematical explanations do yield *counterpossible* information about how things would be different under these impossible scenarios (Baron et al., ming). Efficient coding explanations address w-questions by telling us how things would be different under a range of either counterfactual and counterpossible scenarios. I will now present an extended example of efficient coding explanation in neuroscience, and then discuss its actual and potential applications.

3 LATERAL INHIBITION AND EXPLANATIONS OF EARLY VISUAL RESPONSES

Retinal ganglion cells (RGC's) are the "output" neurons of the mammalian retina. It has long been observed that these neurons have a centre-surround receptive field (RF) organisation. For an ON-centre RGC, when light falls in a certain small, circular area of the visual field, the neuron's rate of firing will increase; and if light falls in the wider area surrounding the centre, then the firing rate will tend to decrease. OFF-centre RGC's have the same concentric receptive field organisation, but with opposite polarity. See Figure 2.

The Difference-of-Gaussian (DoG) function is commonly used to model the RF shape. For an ON-centre cell, the first Gaussian function describes the response of the excitatory centre, with A_1 (height of Gaussian) being the cell's maximum response and σ_1 (spread) describing the spatial extent of the centre. The second Gaussian function, modelling the inhibitory surround, is subtracted from the first. The strength of inhibition is described by A_2, and this takes a lower value than A_1. σ_2 describes the spatial extent of the inhibitory surround, which takes a greater value than σ_1. The DoG model is a two dimensional, circularly symmetrical function in the x, y plane, centred at $(0, 0)$:

$$F(x, y) = \frac{A_1}{2\pi\sigma_1^2} \exp\left(-\frac{x^2 + y^2}{2\sigma_1^2}\right) - \frac{A_2}{2\pi\sigma_2^2} \exp\left(-\frac{x^2 + y^2}{2\sigma_2^2}\right) \quad (1)$$

In his discussion of the DoG function, David Kaplan argues that it is a phenomenological model with high predictive and descriptive value but

lacking explanatory force. Explanations of the neurons' responses, it is argued, will be arrived at once we have modifications of the model which include mechanistic detail:

> Transforming the DOG model ... into an explanatory mechanistic model involves delimiting some of the components and some of the causal dependencies among components in the mechanism responsible for producing the observed structure of the receptive fields, along the lines indicated by 3M. One way to do this, for instance, would be to supplement the model with additional terms corresponding to various components in the retinal ... circuit giving rise to the observed response properties of ganglion ... neurons. (Kaplan, 2011, 360)

Kaplan then references two neuroscientific articles on the retina which proceed in this direction. In contrast with this mechanist perspective on the system, I will discuss a tradition of research which explains the neurons' response properties in terms of the information processing functions which they must perform. This approach proceeds not by adding mechanistic detail to the DoG model but by interpreting it as implementing a particular coding strategy. We should think of the approach as addressing a very different kind of question from the one answered by mechansistic neuroscience—the question of *why* neural systems have the properties that are observed.

The first step is to introduce the concept of *lateral inhibition*. Sensory neurons are said to exhibit lateral inhibition when excitation of one neuron brings about inhibition of the responses of its neighbours. The centre-surround RF's of the retina are indicative of a circuit with lateral inhibition, since the suppressive areas of the RF's arise from the inhibitory inputs of nearby interneurons whose RF's are adjacent in the visual field. Lateral inhibition in the retina is the standard explanation of the visual illusions shown in Figure 3, and it is interesting to note that Ernst Mach posited that the Mach Band illusion was caused by an antagonistic response arrangement in the visual system nearly a century before direct neural recordings were made.

This sounds like the description of a mechanism and one might think that the explanation of Mach bands and the Hering grid would look to be a just a mechanistic one. However, since the 1960's neuroscientists have offered at least three different non-mechanistic explanations for the presence of centre-surround receptive fields and lateral inhibition in the early visual system. These non-mechanistic explanations all refer to the information processing task that has to be performed by the system, and they argue that lateral inhibition serves an important function in the service of this task.

This is a similar contrast to the famous 'how?' vs. 'why?' division in biology. As Barlow (1961b, 782) writes, Ratliff's experiments on the crab's eye "tell us a good deal about *what* the lateral inhibitory mechanism does and something about *how* it does it, but there remains a third question to ask. The fact that this mechanism has evolved independently in a wide variety of sensory relays suggests that it must have considerable survival value: *why* is this so?" Interestingly, this was published the same year as the institutionalistion of the proximate-ultimate distinction by Mayr (1961).

See Ratliff (1965) and **?**. The effects of lateral inhibition have been observed in other perceptual modalities, like touch (von Békésy, 1967, 41-45).

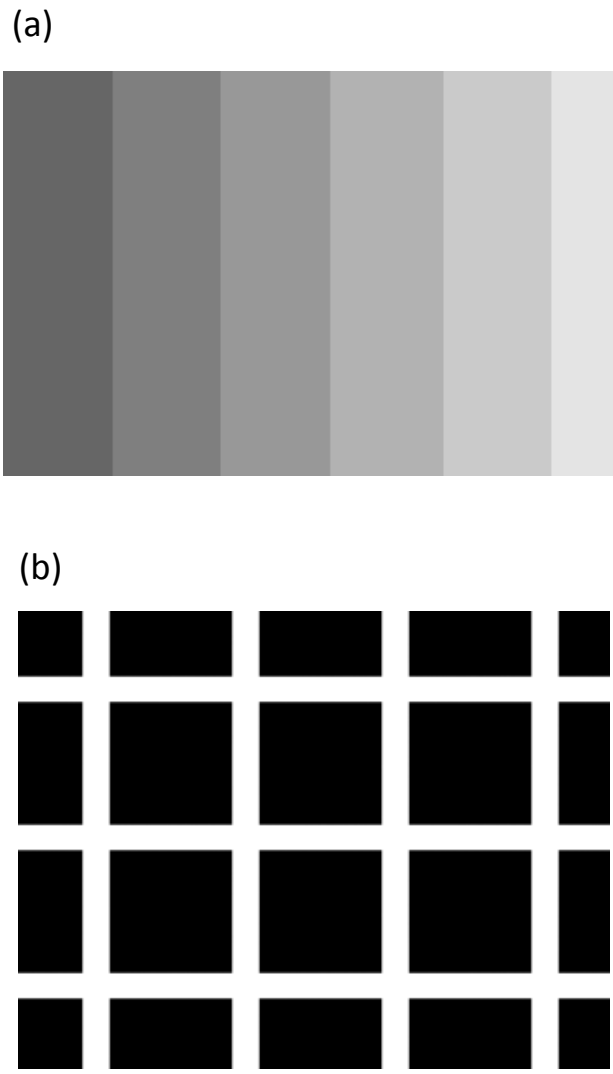

Figure 3: Visual Illusions Explained by Lateral Inhibition. (a) Mach Bands. Within each of the broad vertical bands the grey level is uniform yet we perceive a thin dark vertical strip near the border with a lighter band, and a thin lighter grey strip near the border with a darker band. (b) Hermann Grid. The dark spots at the intersections of the white crosses are illusory. In both cases the illusory patterns are attributed to the presence of inhibitory connections between retinal neurons. As Ratliff (1961, 195) writes regarding the Mach bands, "A unit [whose RF is] within the dimly illuminated region, but near this boundary, will be inhibited not only by dimly illuminated neighbors but also by brightly illuminated ones. The total inhibition exerted on such a unit will be greater, therefore, than that exerted on other dimly illuminated elements that are farther from the boundary; consequently its frequency of response will be less than theirs . . . Thus the differences in activity of elements on either side of the boundary will be exaggerated, and the discontinuity in this pattern of illumination will be accentuated in the pattern of neural response." *Image credit: Wikimedia commons.*

3.1 Edge Detection and Feature Sharpening

All of the efficient coding explanations of lateral inhibition that I will discuss start with the idea that the early visual system must recode the input coming from the photoreceptors and suppress the signals which are not of high value to the downstream visual areas. One can think of the recoding by analogy with image processing routines which reduce the file size of a digital photograph. The data compression can either be "lossy" or "lossless". The first two proposals regarding lateral inhibition differ crucially in where they stand in the "lossiness" of the recoding. The edge detection hypothesis supposes that lateral inhibition serves to detect and/or enhance visual input that is most important to the downstream system—i.e. the edge structure in the visual scene—at the expense of passing on the rest of the input form the receptors. This is a lossy code because non-edge information is supressed by lateral information and this information that is not signalled could not be recovered by the downstream system.

This hypothesis is nicely summarised by Ratliff (1961, 183), one of the neuroscientists who collaborated with Hartline on the seminal research on the *Limulus* eye:

> The interplay of excitatory and inhibitory influences over interconnections within the retina yields patterns of optic-nerve activity that are more than direct copies of the pattern of external stimulation. Certain significant information is selected from the immense detail in the temporal and spatial pattern of illumination on the receptor mosaic, enhanced at the expense of less significant, and only then transmitted to the central nervous system.

Ratliff goes on to say that significant features are edges ("loci of transitions from one intensity to another and from one color to another") and that lateral inhibition in the eye of the *Limulus* is "an integrative neural mechanism which plays a role in the detection and enhancement of such contours."

Alongside edge detection, the notion of feature sharpening or enhancement comes up in Ratliff's discussion of the Mach Band illusion (see Figure 3). As Ratliff (1961, 199-200) also writes,

> These [inhibitory] interactions accentuate contrast at sharp spatial and temporal gradients and discontinuities in the retinal image: borders and contours become "crisp" in their neural representation. Thus, the pattern of optic-nerve activity that results is by no means a direct copy of the pattern of stimulation on the receptor mosaic; certain information of special significance to the organism is accentuated at the expense of less significant information.

Barlow (1961a, 219) calls this the "password hypothesis". In his book on inhibition in various sensory modalities, von Békésy (a Hungarian physicist and physiologist with a background in signal engineering) appears to endorse the proposal that lateral inhibition enables the selective signalling of only important information. Speaking of the explosion of scientific information, he writes that, "[s]urvival requires that we discard the unimportant portions of this information" (von Békésy, 1967, 7). Mach appears as an early proponent of this hypothesis: "Since every retinal point perceives itself, so to speak, as above or below the average of its neighbors, there results a characteristic type of perception. Whatever is near the mean of the surroundings becomes effaced, whatever is above or below is disproportionately brought into prominence. One could say that the retina schematizes and caricatures." Mach (1868), translated in Ratliff (1965, 306)

Other proponents of the edge detection explanation of lateral inhibition are computer vision pioneers, Marr and Hildreth. Rather than beginning with neuroscientific findings, their approach to vision enquires "directly about the information processing problems inherent in the task of vision itself" [Marr and Hildreth (1980, 188); Marr (1982)]. As they see it, the task of the early visual system is to produce, from the raw photoreceptor input, a "primal sketch" of features such as edges, bars and blobs. They show that one way to achieve this is by processing the input image with "Laplacian of Gaussian" filters, mathematical operators which find the areas of steepest illumination change—typically the edges in the image. Their filters are very similar to Difference of Gaussian functions used to model retinal ganglion cell receptive fields, and are identical under certain parameter settings (Marr and Hildreth, 1980, 207, 215-217). So their explanation of centre-surround RF's is that it serves the function of edge detection.

What kind of explanation of lateral inhibition is Ratliff's, on the one hand, and Marr and Hildreth's, on the other? As I see it, we have a case of *functional* explanation. We are told that the function of the early visual circuits which show lateral inhibition and neurons with centre-surround RF's is to detect the edges that are present in the visual scene but are not represented sharply enough in the first encoding at the photoreceptor layer. This fits naturally within a causal framework—the system has the features that it does because it evolved or developed to perform a specific task. The other part of the story is that Marr and Hildreth (1980) present a series of arguments and mathematical proofs to make the case that the image processing steps performed by their Laplacian of Gaussian operator is the optimal way to achieve the required representation of edges. In other words, we have a mathematical and non-causal explanation of why having neurons with the appropriate kind of lateral inhibition (which can be said to implement Marr and Hildreth's operator) is the optimal way to achieve the desired task.

3.2 Redundancy Reduction

The locus classicus for explanations of sensory physiology in terms of redundancy reduction is Horace Barlow's (1961) paper, "Possible principles underlying the transformation of sensory messages." Barlow draws on the influential article by Attneave (1954), which applies Claude Shannon's calculation of the redundancy of written English to the analysis of natural visual stimuli. Information theory provides the mathematical framework for thinking about neural signalling and redundancy. The basic idea is that "sensory relays" (of which retinal ganglion cells are an example) operate to recode information from inputs (ultimately—for RGC's—the photoreceptor layer), in such a way as to economise the consumption of resources (e.g. number of neurons needed, and number of action potentials they fire on average). One way to economise is to reduce the redundancy of the code by eliminating signals which transmit information which is already known or expected by the receiver—see Figure 4. More generally, (Barlow, 1961a, 230) writes, "[t]he principle of recoding is to find what messages are expected

Though as (Barlow, 1961a, 223) notes, the idea was prefigured in the writings of Karl Pearson, Kenneth Craik, Donald MacKay and Ernst Mach.

It bears emphasis that the uptake of information theory from the field of signal engineering to psychology and neuroscience was very rapid. The founding work of information theory (Shannon, 1948), and the paper on redundancy in English (Shannon, 1951) were published only very shortly before Attneave (1954).

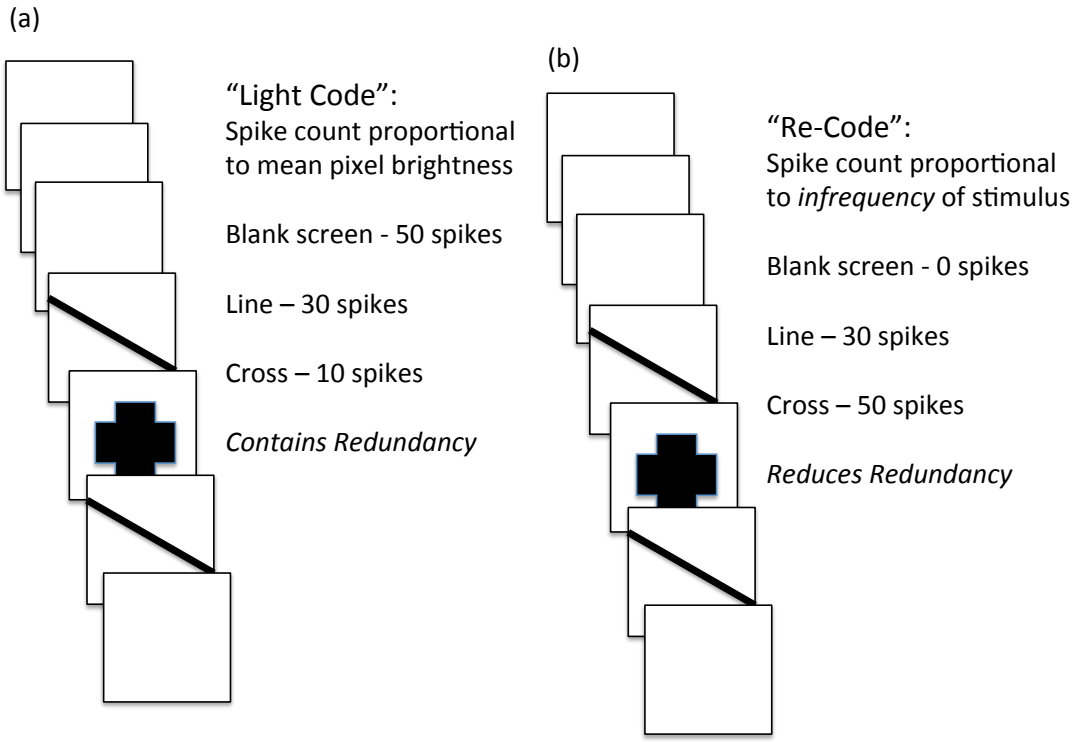

Figure 4: Recoding to Reduce Redundancy. (a) Light Code. Since neural response is proportional to mean pixel brightness, the blank screen will elicit the biggest response. But since the blank screen is frequent, and to be expected by the receiver of the signal, the spikes illicited by the blank screen are redundant. (b) Re-Code. Now the neural response is proportional to the *infrequency* of the stimulus. The blank screen is most frequent, so illicits no response; the cross is most infrequent, so causes the biggest response; and the response caused by the line is intermediate.

on the basis of past experience and then to allot outputs with few impulses to these expected inputs, reserving the outputs with many impulses for the unusual or unexpected inputs."

We can see that the redundancy reducing code in Figure 4(b) is economical or efficient because it uses fewer action potentials to transmit the same amount of information as the first code (a). Since action potential generation is one of the major metabolic costs of the nervous system, it is reasonable to hypothesise that the nervous system, where possible, will operate in such a way as to minimise the number of spikes generate while maintaining the same rate of information transmission. This is how Barlow (1961a, 226) presents the hypothesis:

> We may suppose that the [sensory] relay has a range of possible codes relating input to output: the [redundancy reduction] hypothesis says that, for a given class of input message, it will choose the code that requires the smallest average expenditure of impulses in the output. Or putting it briefly, it economizes impulses; but it is important to realize that it can only do this on the average; the commonly occurring inputs are allotted outputs with few impulses, but there may be infrequent inputs that require more impulses in the output than in the input.

Note that this is a lossless code. The idea is not that the early visual system throws out, or makes unavailable, information that is there in the input concerning the most probable stimuli, but that it does not waste resources in signalling them to downstream receivers.

If we have reason to think that a neural system, like the retina, does indeed implement a redundancy reducing code, then we have an explanation for its observed physiological properties, like the receptive field structure of RGC's. Evidence for the implementation of a particular coding strategy can come in the form of physiological data about the system in question, anatomical findings about circuit structure, and a theoretical arguments that the observed neural system can carry out the computation described by the coding scheme.

In a further article within the same volume as the "Possible Principles" paper, Barlow presents his case that the redundancy reducing hypothesis explains lateral inhibition in the retina. In answer to the question, *why is there lateral inhibition?*, Barlow (1961b, 782) states:

> The suggested answer is that it enables almost the same amount of information to be transmitted with a smaller expenditure of impulses. It is thus an example of a redundancy-reducing code and confers the advantages that Attneave (1954) and I have argued for.

The rest of the article is taken up with a demonstration that lateral inhibition is an effective means of attaining redundancy reduction. This is via an example of the processing of a photographic image in which the brightness value of any small area has the local mean luminance value subtracted from it, thus modelling the effect of inhibition in the retinal system. Barlow observes of the resulting processed image that it retains almost all of the information of the original (the edges and therefore the items in the scene) while needing a much smaller range of brightness values to convey this information. In other words information is compressed but not lost.

We should now consider what kinds of explanation the redundancy reduction hypothesis provides. It strikes me that there are both causal and non-causal dimensions. As apparent in Barlow's discussion of the different explanatory questions (see Footnote above), the redundancy reduction hypothesis is intended to explain what the evolutionary value of lateral inhibition is. Thus the resulting description of the information processing challenge that the retina faces, and the evolutionary pressure towards efficient coding, is a kind of (non-mechanistic) causal explanation. In a very abstract way, it considers environmental conditions and selective pressures, and proposes that lateral inhibition is a result of these factors. For example, we are told that if there were no statistical regularities (spatial or temporal correlations) in natural visual stimuli (in the evolutionary environment of the animal) then the eye could not utilize a redundancy reduced code and we would not expect to see lateral inhibition. On the other hand, Barlow's hypothesis also relies on the mathematical theory of information.

This fits the template of interventionist causal explanation. The redundancy reduction hypothesis tells us that statistical regularities in the visual environment make a difference to the coding schemes employed in the eye. One could perform a practically infeasible, but not modally impossible, experiment where one observes the evolution of creatures in an environment in which the only visual stimuli are random noise—i.e. no spatial or temporal correlations between visual inputs. We would not expect to see the development of lateral inhibition in early visual system. In fact, Barlow's theory would probably predict the atrophy of the visual system, since under these conditions there is literally no visual information provided to the animal and so it cannot use this sensory modality to aid survival.

The laws of information theory constrain the kinds of coding schemes that are efficient, given the actual environment and needs of the animal. In a non-causal sense, information theory 'makes a difference' to the kind of algorithm that the early visual system can implement. What if the laws of information theory were such that the system could reduce redundancy by making spike count proportional to the *frequency* of stimuli? Then you would not expect to have lateral inhibition because it would be efficient for the system to signal mean luminance. There is no way to intervene on laws of information theory, so this experiment is not even hypothetically possible. Yet Barlow's account gives us information about what would happen under such counter-possible scenarios.

For the purposes of this chapter, it is not of paramount importance whether this is a *good* explanation of retinal responses. One theoretical reason for thinking that redundancy reduction is not the only "design principle" which can explain the mammalian retina and other early visual systems is the fact that redundancy reduction trades off against robustness to noise. This is easy to see if we take the example of a telegraph message being sent via an electric cable which experiences random fluctuations in the current or voltage. This noise will result in an error in the decoding of a proportion of the letters sent by the telegrapher. But because of the redundancy within written English, up to a certain percentage of errors it is still quite easy to reconstruct the intended message. In other words, the code is robust to errors introduced due to noise. Since we know that neurons are noisy, this is bound to put constraints on the coding schemes employed by the nervous system. Our last explanation of lateral inhibition does explicitly take noise into consideration.

3.3 Predictive Coding

Unlike the others discussed so far, Srinivasan, Laughlin and Dubs explicitly compare their explanatory hypothesis about the function of lateral inhibition with the alternative proposals. Their claim is that we should think of lateral inhibition as implementing a *predictive code*, and that this account

---

For evidence that the retina does not always follow a redundancy reducing strategy because it fails to decorrelate the responses of neighbouring RGC's, see Puchalla et al. (2005) but also Doi et al. (2012). Barlow (2001) presents an extensive and deep criticism of his 1961 redundancy reduction argument.

E.g. it is possible to infer that the message corrupted with an error rate of about 0.25, "Tve Uing is Dqad, Lobg Yive bhe Queec", means "The Kind is Dead, Long Live the Queen" because of the redundancy of written English—i.e. "the amount of constraint imposed on a text in the language due to its statistical structure, e.g., in English the high frequency of the letter E, the strong tendency of H to follow T or of U to follow Q" (Shannon, 1951, 50). A telegraphic code which removed redundancy would not transmit the letters that can be predicted from our knowledge of the structure of English, such as an 'h' following a 't' in the words 'the' (just as the redundancy reducing code in Figure 4(b) does not send a spike when it encounters the most likely stimulus pattern. However, if this streamlined code were to be corrupted by noise it would not be decodable.

(Puchalla et al., 2005, 501) write that, "While there has been a great focus on efficiency as a fundamental design principle for neural codes, robustness is less well understood .... Quantifying our intuitive notion of robustness, ..., promises to enrich our understanding of design principles in neural networks. Especially interesting will be to explore how redundancy and efficiency trade off as the signal-to-noise ratio of visual stimuli changes."

There has been much discussion in recent philosophy of mind of the proposal that predictive coding provides a single unified framework for thinking about neural processing and cognitive function. See Hohwy (2013) and Clark (2016). Note that the proposal here is much more modest in that it only extends to one specific circuit, and much more concrete in that it tells us exactly how the predictive code could be implemented by the circuit in question. One

subsumes both the edge detection and redundancy reduction proposals. These authors are neuroscientists whose research focuses on the visual system of the fly which, like the horseshoe crab, has a compound eye, and has neurons (large monopolar cells) with circular surround RF's .

The predictive coding hypothesis is stated as follows:

> The antagonistic surround [inhibitory area of the RF] takes a weighted mean of the signals in neighbouring receptors to generate a statistical prediction of the signal at the centre. The predicted value is subtracted from the actual centre signal, thus minimizing the range of outputs transmitted by the centre. In this way the entire dynamic range of the interneuron can be devoted to encoding a small range of intensities, thus rendering fine detail detectable against intrinsic noise injected at later stages in processing. This predictive encoding scheme also reduces spatial redundancy, thereby enabling the array of interneurons to transmit a larger number of distinguishable images, taking into account the expected structure of the visual world. (Srinivasan et al., 1982, 427)

The idea is that the surround portion of the neuron's receptive field measures local mean luminance, giving a prediction of what the luminance will be in the centre. If this prediction is accurate, then the luminance value at the centre will be exactly cancelled out by the inhibitory input to the centre, and the cell's firing will not increase. But if the central luminance value diverges from the prediction, then it will overcome the inhibition and a signal will be generated to say that something "surprising" is happening in the centre. Unlike (Barlow, 1961a, 224), they also emphasise that lateral inhibition, understood in their way, has advantages for systems like actual neural ones, which have high intrinsic noise.

Srinivasan et al. (1982, 428) point out that the predictive coding idea first came from television engineering, citing papers by Oliver (1952) and Harrison (1952). The predictive coding hypothesis has recently been employed by Sterling and Laughlin (2015, 249) in their comparison of early visual processing in mammals and flies. They write that, "predictive coding, an image compression algorithm invented by engineers almost 60 years ago to code TV signals efficiently, is implemented in animals by a basic sensory interaction". Again, the idea is that we formulate an explanation of why the neural circuit has an observed feature by showing that it implements an algorithm known to be efficient—both in biological and artificial systems.

---

interesting point of comparison is that the recent philosophical literature does not, to my knowledge, discuss efficiency arguments for predictive coding.

E.g. Srinivasan et al. (1982, 451) write: "in common with alternative functions of lateral inhibition, edge detection and predictive coding are in no way exclusive. The more advantages a given filtering or sampling procedure has, the better! The difference is that predictive coding takes into account the qualities of the retinal image in order that it might be encoded within the constraints imposed by neuronal signals. By comparison, edge detection isolates a single characteristic of a scene, that can, through its spatial distribution, provide an adequate and compact description, thought suitable for subsequent processing at higher levels."

"Interneurons exhibiting centre-surround antagonism within their receptive fields are commonly found in peripheral visual pathways. We propose that this organization enables the visual system to encode spatial detail in a manner that minimizes the deleterious effects of intrinsic noise, by exploiting the spatial correlation that exists within natural scenes." (Srinivasan et al., 1982, 427)

Note that Sterling and Laughlin (2015) are not claiming that predictive coding can explain *all* observed features of the retina or fly's "lamina". Their analysis goes into much detail about different anatomical and physiological features at each layer of the retina and lamina, and presents various efficient coding arguments to explain those observations.

As in the previous two examples, there are both causal and non-causal features to this explanation. Sterling and Laughlin (2015, 249) place much emphasis on the tight energy budget of the central nervous system. This is a causal explanation of neural design, which tells us that if the energy budget were more ample, or if spikes cost fewer molecules of ATP, then we could expect different circuits. Alongside this reasoning, there is the mathematical argument that predictive coding is an efficient means to transmit visual information. This reasoning explains why a neural circuit for visual signal transmission, with a tight energy budget, would be constrained to implement predictive coding through lateral inhibition.

3.4 Some Observations

Before moving on, I would like to say a few words about what we have learned from this case study. I have sketched a historical narrative of the development of contrasting efficient coding explanations of one neural phenomenon, lateral inhibition, in order to make the case that this approach has been an active area of research, alongside the mechanistic one, since the very beginnings of physiological investigation of the visual system. In other words, as soon as neuroscientists were able to measure the effects of visual stimulation on specific neurons in the early visual system, and plot their receptive fields, they began theorising about the functions of those RF's and discussing abstract coding schemes which could be said to be implemented by the neural circuit. Researchers taking this approach have been very much in the mainstream of visual neuroscience.

The other point I would like to make here is that in each of the cases presented above, ideas about what the visual system was coding, and why, have been inspired quite directly by work outside of neuroscience: information theory and signal engineering, computer vision and television engineering. Do origins of the efficient coding approach in engineering shape the practical applications of its findings? How are the reverse engineering of the brain and the forward engineering of brain-like machines connected?

4 PUTTING EFFICIENT CODING EXPLANATIONS TO USE

4.1 Scaling Data Mountain

Neuroscience does not suffer from a poverty of data. According to Hill (2015, 113), the rate of publication in neuroscience has grown from 30,000 articles per year in 1990 to 100,000 per year in 2013. What's missing is the means for neuroscientists to streamline and consolidate the deluge of results so that it is clear to each subfield what is known and what is not known.

At the beginning of their recently published book on the efficient coding approach to neural systems, Sterling and Laughlin (2015) are clear that they see their work as offering ways to digest the surfeit of data—or to switch to their metaphor, to climb the mountain of data. Their strategy is to articulate a small number of "organizing principles" that afford efficient coding explanation of diverse features of biological information processing in organisms spanning the chain of being, including bacteria,

flies, and human brains. Many of these "design principles" come directly from engineering and information theory, while others are based on direct measurement of the cost of information processing in biological tissue. The basic idea is that by focussing on the information processing *function* of neural systems, scientists will be better able to discern the really important phenomena against the background of extraneous mechanistic detail.

Interestingly, this motivation for the efficient coding approach was already stated by (Barlow, 1961a, 217).

> A wing would be a most mystifying structure if one did not know that birds flew. ...[W]ithout understanding something of the principles of flight, a more detailed examination of the wing itself would probably be unrewarding. I think that we may be at an analogous point in our understanding of the sensory side of the central nervous system. We have got our first batch of facts from the anatomical, neurophysiological, and psychophysical study of sensation and perception, and now we need ideas about what operations are performed by the various structures we have examined. ...
>
> It seems to me vitally important to have in mind possible answers to this question when investigating these structures, for if one does not one will get lost in a mass of irrelevant detail and fail to make the crucial observations.

From our study of lateral inhibition we can already see how efficient coding explanations can be used to streamline and consolidate neuroscientific facts. As pointed out above, the eyes of mammals, crustaceans and insects vary quite considerably in their anatomical and physiological details. By focussing on the *what?* and *how?* questions one could get lost in the mechanistic detail of each eye's neural circuit: the layout of the neurons, their dendritic arbors and activity patterns. In contrast, if one focusses on the question of *why* the neurons of a particular eye form an inhibitory network, and formulates an efficient coding explanation, the mechanistic details recede to the background and the similarities across mechanistically diverse systems become apparent. The key explanandum phenomenon is the kind of information processing that the inhibitory network affords, and since the explanans is an abstract coding scheme we need not worry too much about the details of biological implementation in each case (so long as a proposed implementation is not inconsistent with the known data).

This has echoes of the idea that explanation proceeds by showing that a set of seemingly unrelated phenomena can be unified with the same explanatory model or theory (Kitcher, 1981). In fact, this remark by Hempel on explanation and unification is very much of a piece with Sterling and Laughlin's stated aims:

They list ten such principles: "compute with chemistry; compute directly with analog primitivies; combine analog and pulsatile processing; sparsify; send only what is needed; send at the lowest acceptable rate; minimize wire; make neural components irreducibly small; complicate; adapt, match, learn, and forget." (Sterling and Laughlin, 2015, ii)
This sentiment is echoed by (Marcus and Freeman, 2015, xii), quoted at the start of Section 4.3. As it happens, one ongoing project in retinal anatomy that has received much attention (and criticism) is Sebastian Seung's crowdsourcing challenge to get the complete wiring diagram (*connectome*) of the mouse retina. Much criticism has focussed on the issue that there is so much difference in the detailed anatomy even amongst individuals of the same species, that a dense reconstruction of the wiring cannot be practically or theoretically informative. But see Kim et al. (2014).

> What scientific explanation, especially theoretical explanation, aims at is not [an] intuitive and highly subjective kind of understanding, but an objective kind of insight that is achieved by a systematic unification, by exhibiting the phenomena as manifestations of common, underlying structures and processes that conform to specific, testable, basic principles. Hempel (1966, 83), quoted by Kitcher (1981, 508).

I should note, however, that Sterling and Laughlin's declared inspiration is not twentieth century philosophy of science but the unsurpassed subsumption of disparate data under unifying theory that was afforded by the theory of natural selection (Sterling and Laughlin, 2015, xiv). Moreover, the explanatory sufficiency of efficient coding reasoning does not thereby stand and fall with the covering law and unificationist model of explanation. As I have been careful to point out, efficient coding explanations satisfy the requirement of answering w-questions which many critics of covering-law explanation subscribe to.

4.2 Forward Engineering

Sterling and Laughlin's goal is to reverse engineer the brain. They do not discuss ways that the efficient coding approach could be applied beyond basic neuroscience, in neuro-inspired technologies and bio-engineering involving the brain. However, this is an increasingly active field of research and it is interesting to see how efficient coding explanations play a role in it.

More specifically, the concepts of efficient coding explanation—e.g. constraints, trade-offs, efficiency, redundancy and optimisation—come ultimately from engineering. While computational neuroscientists are taking a design stance to neuro/bio systems and doing the reverse engineering, the principles that they formulate or discover (see footnote) will often apply equally to man-made systems and biological ones. This is necessarily the case when the principle in question is a result derived from information theory or any kind of mathematical or statistical argument. The trade-offs revealed by the mathematical analysis of information transmission can be thought of as design constraints that an information engineer ought to be conscious of, and knowledge of biological "solutions" frequently inspires better design. So even when trade-offs, such as the one between redundancy and robustness, cannot themselves be subject to intervention, knowledge of those trade-offs can has very direct practical application.

One of the motivations for studying the coding schemes which allow the brain to process information with much less power consumption than artificial computers is in order to design computers which are themselves more efficient. Rahul Sarpeshkar, whose hybrid coding argument was discussed above, is himself an electronics engineer with a research focus on low-power biology-inspired computation. For example, his ideas have applications in the design of implantable medical electronics such as sensory-substitution devices (Sarpeshkar, 2010).

In the field of vision science we see the influence running from engineering to neuroscience and back again. We saw in our case study of lateral inhibition, neuroscientists borrowed concepts from signal engineering and

There is a parallel here with work in *synthetic biology*. See Knuuttila and Loettgers (2013a) and Knuuttila and Loettgers (2013b)

information theory in order to explain their observations. From the 1970's onwards there have been concerted efforts to design algorithms which will give computers or robots functioning vision. Though Marr (1982) famously argued that computer vision research was best off proceeding independently of visual neuroscience, bracketing questions about neural implementation, I think we should understand this as a warning against focussing on irrelevant mechanistic issues. For in Marr and Hildreth (1980) much attention is paid to the comparison between their Laplacian of Gaussian filter and empirical findings in psychology and neuroscience about the workings of the early visual system, where these findings are concern the abstract coding schemes employed here rather than detailed anatomy or physiology.

Another example is the use of the Gabor function to model of neurons in primary visual cortex [see Chirimuuta (2014, §5.2) and Chirimuuta (forthcoming, §3)]. The introduction of the function, borrowed from mid-twentieth century communications engineering, was justified by Daugman (1985) as the optimal solution to the joint problem of decoding both spatial location and spatial frequency (width of edge) information. John Daugman is himself a computer scientist who has sought to design better image recognition algorithms on the basis of his study of visual cortex.

Furthermore, the engineering approach can also be applied to the manipulation of the brain itself, not just in the building of artificial devices. Neuro-engineering is a fast growing field of activity involving the development of *brain computer interfaces* (BCI's) which read off and decode neural activity in order to control external devices such as computers and robotic limbs, or to channel information directly into the brain. In order for such technologies to be effective, the brain's activity must be understood in abstract enough terms to allow for translation to and from digital computers. That is, the "neural code"—the information conveyed by particular patterns of activity—must be deciphered and manipulated in a way that is independent of the specific biological implementation (Chirimuuta, 2013). This is why abstraction from mechanistic details, and recourse to rarefied mathematical descriptions of signals is particularly useful here. Yet in order to build an effective BCI, a brilliant decoding algorithm is not enough. One also needs an electrode implant in the cortex which has long term stability and does not quickly lead to degeneration of the neural tissue in which it is embedded. Of course this requires precise anatomical knowledge of the cortical layers, knowledge of the biochemical environment, and of neural cell death cascades—in other words, a detailed mechanistic understanding of the brain. In short, this is a field of endeavour in which mechanistic and efficient coding knowledge are both integral to its success.

4.3 Defining Neural Computation

It is uncontroversial, amongst neuroscientists, to say that the brain computes (Koch, 1998, 1). And it is by now well established that the brain does not compute in the same way that a general purpose digital computer does, or in the fashion of any known analogue machine. I concur with Piccinini and Bahar (2013, 476) that neural computation is sui generis. The tricky

Note also that computer vision algorithms which employ lateral inhibition—e.g. by using the DoG function—are quite commonly used. See Klette (2014, 75-76), Moini (2000, 18-19), and Lyon (2014) on the invention of the optical mouse.

thing is then to put some useful definitions in place which will help clarify what is or should be meant by neural computation, and there is not yet a consensus emerging from the discipline of theoretical neuroscience. As Marcus and Freeman (2015, xii) write, "we have yet to discover many of the organizing principles that govern all that complexity. We don't know, for example, if the brain uses anything as systematic as, say, the widespread ASCII encoding scheme that computers use for encoding words. And we are shaky on fundamentals like how the brain stores memories and sequences events over time."

Piccinini and Bahar (2013, 477-9) assert that computation is a kind of "mechanistic process", and thus that the empirical study of neural mechanisms, and the search for mechanistic explanations of the brain and psychological states, will eventually lead to an understanding of neural computation. I believe that this approach is misguided. As we saw in the case study of lateral inhibition, any restricted focus on the mechanistic details giving rise to inhibitory effects would not be illuminating as to the computational properties of the circuit. For one thing, the search for mechanistic explanations does not draw from the theoretical frameworks in engineering and mathematics which can be used to characterise computational systems. For another, the mechanistic perspective obscures the interesting commonalities amongst biophysically very different systems. It was only by taking the efficient coding perspective, and asking in abstract terms what function the circuit performs and why, that hypotheses could be formed about what coding scheme is implemented in these systems.

In order to make progress towards a definition and theory of neural computation, general coding schemes and unifying principles are far more valuable than a disunified collection of data concerning mechanisms in the brains of different animals. This requires that scientists work with a "level of description" which is abstracted from that of mechanistic implementation [cf. Marr (1982); Carandini (2012)], and is assumed in the efficient coding tradition. One idea along these lines which has recently been attracting attention is that of *canonical neural computations* (Carandini and Heeger, 2012). These are computational operations which are frequently used to model small circuits and are found to reoccur in different species and brain regions. The DoG model of lateral inhibition would be an example, and they are commonly invoked in efficient coding explanations (Chirimuuta, 2014). Carandini and Heeger's proposal is to identify a handful of such computations which might be thought of as the building blocks for more complex neural computations. If the project is successful, the result will be a clearly articulated theory of neural computation.

5 CONCLUSION

In this paper I have charted the development of efficient coding explanations of a well known neural phenomenon, and discussed practical applications of these and other models and explanations. I have been somewhat diffident about the causal/non-causal distinction because in practice these aspects of efficient coding explanation are integrated and complementary to one another. What is more significant is the difference between efficient coding and mechanistic explanation, since each approach reveals and obscures

But see Koch (1998) for a hybrid computatational-mechanistic approach.

different aspects of a neural system. For example, efficient coding models tend to mask the bio-chemical intricacy of the brain's 'circuits', treating them more like arrays of electronic switches. As a result, such models do not play a role in the development of pharmaceuticals to alleviate organic diseases affecting brain cells; they do make a difference, however, in the design of prosthetic systems which aim to replace lost neural tissue. More generally, they have an important place in tasks where 'big picture' ideas about the system's function are needed.

Throughout this paper I have emphasised the extent to which the efficient coding framework draws from the theories and concepts of communication engineering. I would like to finish with the caveat that this analogical approach to understanding the brain brings with it its own limitations. Both neuroscientists and philosophers of neuroscience should be aware of the ways in which the analogy between the brain and a man-made computer or signalling system can break down. As (Barlow, 2001, 244) puts it, "[i]n neuroscience one must be cautious about using Shannon's formulation of the role of statistical regularities, because the brain uses information in different ways from those common in communication engineering." The challenge is to find out exactly *how* the brain uses information, and what "information" is in the context neuroscience rather than engineering. The efficient coding approach is just a starting point.

REFERENCES

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

Attneave, F. (1954). Some informational aspects of visual perception. *Psych. Rev. 61*, 183–193.

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.

Barlow, H. (2001). Redundancy reduction revisited. *Network 12*, 241–253.

Barlow, H. B. (1961a). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory Communication*, pp. 217–234. Cambridge, MA: MIT Press.

Barlow, H. B. (1961b). Three points about lateral inhibition. In W. A. Rosenblith (Ed.), *Sensory Communication*, pp. 782–786. Cambridge, MA: MIT Press.

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

Bechtel, W. and R. Richardson (2010). *Discovering Complexity*. Cambridge, MA: MIT Press.

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.

Carandini, M. (2012). From circuits to behavior: a bridge too far? *Nature 15*(4), 507–509.

Carandini, M. and D. J. Heeger (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience 13*, 51–62.

Chirimuuta, M. (2013). Extending, changing, and explaining the brain. *Biology Philosophy 28*(4), 613–638.

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

Chirimuuta, M. (forthcoming). Explanation in neuroscience: Causal and non-causal. *British Journal for Philosophy of Science*.

Clark, A. (2016). *Surfing Uncertainty*. Oxford: Oxford University Press.

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.

Doi, E., J. L. Gauthier, G. D. Field, J. Shlens, A. Sher, M. Greschner, T. A. Machado, L. H. Jepson, K. Mathieson, D. E. Gunning, A. M. Litke, L. Paninski, E. J. Chichilnisky, and E. P. Simoncelli (2012). Efficient coding of spatial information in the primate retina. *The Journal of Neuroscience 32*(46), 16256 –16264.

Hales, T. C. (2001). The honeycomb conjecture. *Discrete Comput Geom 25*, 1–22.

Harrison, C. W. (1952). Expriments with linear prediction in television. *Bell Syst. tech. J. 31*, 764–783.

Hempel, C. G. (1966). *Philosophy of Natural Science*. Englewood Cliffs: Prentice-Hall.

Hill, S. (2015). Whole brain simulation. In G. Marcus and J. Freeman (Eds.), *The future of the brain*, pp. 111–124. Princeton, NJ: Princeton University Press.

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

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.

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

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.

Kim, J. S., M. J. Greene, A. Zlateski, K. Lee, M. Richardson, S. C. Turaga, M. Purcaro, M. Balkam, A. Robinson, B. F. Behabadi, M. Campos, W. Denk, H. S. Seung, and the EyeWirers (2014). Space–time wiring specificity supports direction selectivity in the retina. *Nature 509*, 331–336.

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science 48*(4), 507–531.

Klette, R. (2014). *Concise Computer Vision*. London: Springer.

Knuuttila, T. and A. Loettgers (2013a). Basic science through engineering? synthetic modeling and the idea of biology-inspired engineering. *Studies in History and Philosophy of Science, Part C 48*, 158–169.

Knuuttila, T. and A. Loettgers (2013b). Synthetic modeling and mechanistic account: Material recombination and beyond. *Philosophy of Science 80*(5), 874–885.

Koch, C. (1998). *Biophysics of Computation: Information Processing in Single Neurons*. New York: Oxford University Press.

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

Lyon, R. F. (2014). The optical mouse: Early biomimetic embedded vision. In B. Kisačanin and M. Gelautz (Eds.), *Advances in Embedded Computer Vision*. Heidelberg: Springer.

Mach, E. (1868). über die physiologische wirkung räumlich vertheilter lichtreize (vierte abhandlung). *Sitzungsberichte der mathematisch-naturwissenschaftlichen Classe der Kaiserlichen Akademie der Wissenschaften 57*(2), 11–19.

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

Marcus, G. and J. Freeman (2015). Preface. In *The Future of the Brain*, pp. xi–xiii. Princeton, NJ: Princeton University Press.

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

Marr, D. and E. Hildreth (1980). Theory of edge detection. *Proc. R. So. Lond. B 207*, 187–218.

Mayr, E. (1961). Cause and effect in biology. *Science 134*, 1501–1506.

Moini, A. (2000). *Vision Chips*. Dordrecht: Kluwer.

Oliver, B. M. (1952). Efficient coding. *Bell Syst. tech. J. 31*(724-750).

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

Puchalla, J., E. Schneidman, R. Harris, and M. J. Berry (2005). Redundancy in the population code of the retina. *Neuron 46*, 493–504.

Ratliff, F. (1961). Inhibitory interaction and the detection and enhancement of contours. In W. A. Rosenblith (Ed.), *Sensory Communication*, pp. 183–203. Cambridge, MA: MIT Press.

Ratliff, F. (1965). *Mach Bands: Quantitative studies on neural networks in the retina*. San Francisco: Holden Day-Inc.

Reutlinger, A. (2012). Getting rid of interventions. *Studies in History and Philosophy of Biological and Biomedical Sciences 43*, 787–795.

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.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal 27*, 379–423; 623–656.

Shannon, C. E. (1951). Prediction and entropy of printed english. *Bell Syst. tech. J. 30*, 50–64.

Sprevak, M. (2012). Three challenges to chalmers on computational implementation. *Journal of Cognitive Science 13*, 107–143.

Srinivasan, M., S. Laughlin, and A. Dubs (1982). Predictive coding: a fresh view of inhibition in the retina. *Proc Biol Sci 216*, 427– 459.

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

von Békésy, G. (1967). *Sensory Inhibition*. Princeton, NJ: Princeton University Press.

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

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.