Computational models of visual processing: a discussion of normalization and predictive coding using Marr's three levels of analysis

Barrie Klein 3386686

Abstract

Our percept is the result of processing visual information by our visual system. This system contains many billions of neurons and a manyfold of connections between these neurons. In order to understand a complex system like our visual system, it has been proposed to approach it at different levels of analysis. This proposal distinguishes three levels: the first level describes the computational goal of the system, the second level the algorithms the system uses to achieve this goal, and the third level describes the way in which these algorithms are physically implemented in the brain. Focusing mainly on the first two levels of analysis, this review describes two computational models: the normalization model and the predictive coding model. This discussion shows that, despite little is known about the biological mechanisms underlying the models, the behavior of the visual system can be very well understood when computational models at the remaining two levels of analysis are applied to its behavior. Furthermore, this discussion shows that the models' algorithms may supplement one another at certain stages of visual processing. Finally, this review may provide support for a relation between predictive coding and normalization models. Several implications of this possible relation are discussed.

Introduction

Vision is the dominant sense in humans and most of our daily interactions require people to have full vision[1]. In our brain, about 25% of the cerebral cortex is devoted to processing visual information. Our visual percept is the output of a complex system that includes billions of neurons and many more connections between them. In order to understand complex systems such as the human visual system, Marr and Tomaso have proposed three levels of analysis[2]. At the first level (computational) the goal of the system is described. More specifically, one determines what problems the system has to overcome. The second level (algorithmic), describes how the system does what it does. In other words, what representations does the system use and what algorithms are used to modify these representations. The third level (physical) is concerned with how the system is physically realized, i.e. what neural structures and activities 'carry out' the algorithms considered in the level above? Generally, computational models explaining properties of visual processing can be described at one or more of these levels.

Using Marr’s levels of analysis, I will discuss two computational models and will mainly focus on their computational goals and algorithmic implementations. The first model is the normalization model proposed by Heeger[3]. The computational goal of this model is to deal with non-linearities of neural responses. The second model discussed is predictive

coding[4]. This model postulates that the brain uses the predictability of visual information to optimize visual processing. Depending on the framework from which it is considered, it can serve the goal of dealing with uncertainty inherent to visual perception (Bayesian brain framework[5,6]), or the goal of improving efficiency of visual processing (information theoretic approach[7-9]).

For both models, I will first explain the problems with which the models deal (computational goal) and the algorithms proposed to achieve this goal. Then, studies that support the employment of these algorithms by our visual system will be discussed. This discussion demonstrates that the proposed algorithms capture a variety of properties of the visual system, suggesting that the brain employs similar algorithms. Furthermore, this demonstrates that although little is known about the biological mechanisms that implement the proposed algorithms (Marr's physical level), one can gain an understanding of the visual system by attempting to explain phenomena at a higher level of analysis. Finally, despite the apparent dissimilarity of these models at the computational and algorithm level, the possibility and potential implications of normalization and predictive coding capturing similar physical mechanisms will be discussed.

Normalization algorithms deal with non-linearities in neural responses

Heeger[3] proposed the normalization model as a potential mechanism underlying a neuron's non-linear response. Non-linear in this case means that the response intensity of a neuron (e.g. the amount of spikes per second) is not just a simple reflection of its input intensity. If a neuron would respond in a linear way, each increase in input intensity is accompanied by a roughly fixed increase in the output response. This can be likened to a scale. Assuming you are using one that functions properly, the more weight you put on the scale, the higher the weight indicated by the scale. The indication by the scale is a direct function of the weight on top of the scale. The dashed line in figure 1 shows a simple linear relationship you would find between the weight placed on top of the scale and the weight indicated by the scale. This line, however, is rather different from the true relationship measured between a neuron's input and output, which is given by the solid line in figure 1. In terms of our scale analogue, it appears that our scale does not reflect reliably weight increases below and above a certain weight. To account for this non-linear relationship between a neuron's input and response, it has been suggested that neighboring neurons mutually inhibit each other's responses[10,11]. In terms of our scale, its weight indicator would be impeded by weights placed on other scales that can somehow influence our scale. Heeger[3] formalized this property of neurons in a mathematical algorithm, which is known as normalization. It assumes that the response modification of a neuron can be captured by dividing its response by a measure of the activity of neurons that inhibit it. When applied to our scale example, one could say that the initial indication of a weight on our scale is divided by a number that represents the indication of weights on other scales. This number does not necessarily have to be a sum of these indications, but could represent other measures as well. It is important to note that, although mutual inhibition between neurons is the underlying mechanism behind normalization, this does not say anything about the exact connections between neurons that support this behavior of neurons,

i.e. this is not an exhaustive physical explanation of normalization. In that sense, the activity of surrounding neurons is used to determine the value of one of the term in the normalization algorithm, namely the divisive term.

Normalization is an extension of an early view of neurons in which they are regarded as linear operators[3,12,13]. To understand this, one has to realize that a neuron receives input from multiple lower level neurons or, in the case of the visual system, photoreceptors. A linear operator would just sum all of its inputs and its response would be proportional to this sum. Before summing the inputs, a linear operator can weigh its inputs. This can be regarded as a process that determines the importance of each input in determining the neuron's response. The higher an input's weight, the more important it is in determining the sum of the inputs and subsequently the response of the neuron. This summing and weighing of inputs results in a measure of a neuron's response non-normalized response. Our scale analogy is a good example here. The weight indicated by the scale is the sum of the weight of all the objects placed on top of the scale. The only oddity about this 'neural scale' is that not every location of the scale's top is treated equally, some parts influence the indicated weight more than other parts of the scale's top. A mathematical form of this process is given by equation 1:

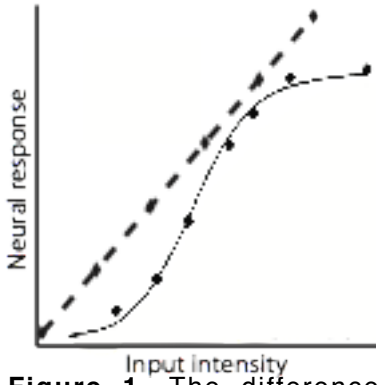

Figure 1. The difference between linear increments of a response (dashed line) and measured neural response increments (solid line) as a function of stimulus intensity. Measured data from ref. 16.

Equation 1: $$D = \sum w_{jk} I_k$$

D represents the response of the neuron *j* (the weight indicated). W_{jk} represents the weights neuron *j* assigns to each of its inputs and this can be regarded as a row of numbers (vector) ranging between 0 and 1 (the unequal treatment of locations across the scale's top). I_k represents the inputs to neuron *J* (the objects placed on top of the scale). This input can be considered a row of numbers with the same length as W_{jk}. All the values represented by I_k are multiplied by those represented by W_{jk} and values of the resulting vector are summed to result in just 1 value, which represents the non-normalized response of the neuron: D. In normalization, this measure of the neural response is divided by a measure of activity of the neurons that inhibit it. Equation 2 describes this computation:

Equation 2: $$R_j = \gamma \frac{D_j^n}{\sigma^n + \sum_k D_k^n}$$

D_j stands for the unnormalized response of the neuron. D_k represents the activities of the neurons that inhibit the neuron. This measure is obtained in the same way as the activity of neuron itself is obtained. To their summed activity the constant σ is added (a number that is added to the measure of the inhibitory scale's activity). This term prevents division by zero

and determines the input intensity at which the neuron reaches its half-maximum response intensity. The lower this term, the sooner the neuron reaches this point and the sooner it reaches its maximum response intensity. Changes to his term have the same effect as changes in the activity from the normalization pool: it shifts the position of the function that describes the relationship between the input intensity and output response of the neuron on the x-axis (figure 2A). γ represents the maximal attainable response intensity of the neuron and is a measure of overall responsiveness of the neuron (the maximal weight the scale can indicate). The higher this term, the higher the response of the neuron can get. The exponent *n* determines the steepness of the neuron's response function (figure 2B), which means that it determines the size of the response increment in relation to an input intensity increment (how much does the indicated weight increase with more weight placed on top of the scale). The higher this value, the steeper the response function and the smaller the range of input intensities to which a neuron responds[14].

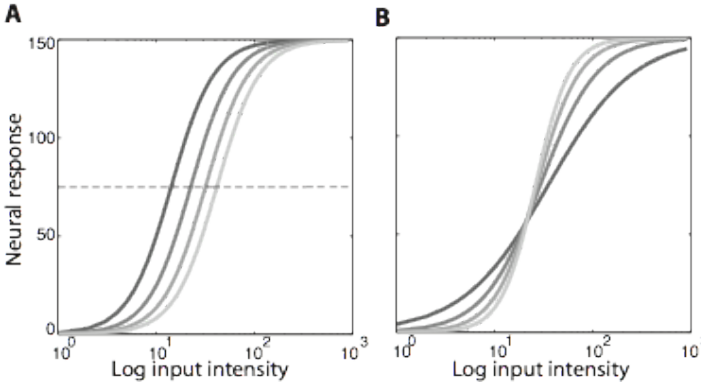

Figure 2. The effect of increasing either σ or the summed activity of the normalization pool from equation 2. The dashed line marks the half-maximum response intensity and γ was set to 150. Lighter curves represent higher values for these terms (**A**). The effect of changing the exponent n. Lighter curves represent a higher value for this term (**B**).

Thus within normalization, a neuron has a receptive field (a scale's top), which is represented by the summation over its inputs and a normalization field, a pool of neurons from which it receives inhibition. The receptive field determines to which stimuli in which location of the visual field the neuron responds. In other words, it determines the spatial and feature tuning of a neuron (the scale analogy is a bit awkward here, but applied to it, this means that not every object's weight is treated equally). The normalization field is usually larger than the receptive field and is broader tuned to features. This means that although stimuli are placed at a location that does not drive the neuron, they can nevertheless inhibit a neuron or when a stimulus contains a feature (for example has a certain orientation) to which the neuron is not tuned, it can nevertheless inhibit the neuron.

Shortly, normalization can be regarded as an algorithm whose computational goal it is to explain non-linearities in neural responses. It does this by taking into account the response of neighboring neurons and it postulates that the normalized neural response can be obtained by dividing the non-normalized, linear response by the activity of the normalization pool. It thereby assumes that responses of nearby neurons are an important parts of its biological implementation. The normalization pool is usually broadly tuned and represents responses of neurons nearby in both the spatial and feature dimension. It has to be pointed out that normalization is very similar to gain control. In which gain stands for the responsiveness of a neuron (i.e. the relation between input and response) which is controlled by some population

of neurons. Normalization is a special case of gain control, in which the signals that control the gain include the signal that determine the neuron's response, i.e. the neuron's own activity contributes to the activity of its normalization pool[14]. The normalization algorithm is applied to many stages along the visual processing stream. For example to light adaptation in the retina[15,16], contrast saturation, size-tuning and masking effects in LGN[17], contrast saturation, surround-suppression, cross-orientation suppression and several other properties of neurons in V1[14], motion processing in MT[18], and even multi sensory integration[19]. This diversity of phenomena captured by normalization strongly suggests that its algorithm represents a computational process that the brain uses across a large variety of processing mechanisms.

Normalization algorithm applied to response properties across the visual system

Light adaptation

As noted above, normalization has been applied to a variety of processes. One of these is light adaptation. The range of light intensities that reach our retina is far greater than the range of light intensities that our visual system can reliably distinguish[20,21]. Therefore, a process is needed that adjusts the sensitivity of our visual system, depending on the overall light intensity. It has been found that the sensitivity of photoreceptors is adjusted depending on the intensity of background light[15,16]. This effect can be captured by a normalization algorithm in which the response of an individual photoreceptor is divided by the response of surrounding receptors, which is mathematically similar to equation 2. Increasing the denominator in this equation (in this case, increasing the measure of the local light intensity) results in a shift of a photoreceptor's response function towards the right, which captures the decrease in a photoreceptor's sensitivity with increasing background intensity nicely (as illustrated in figure 2A). Due to this shift of the photoreceptor's response function, the photoreceptor discounts the local mean light intensity from its response[14]. This effectively results in the photoreceptor signaling deviations from the local mean light intensity, which is actually a measure of local contrast. This is another property of visual scenes whose processing shows a number of non-linearities that can be captured with a normalization approach as well.

Non-linearities in contrast processing

Bonin and his colleagues[17] proposed a model to account for three properties of contrast responses of neurons in the lateral geniculate nucleus (LGN). First, responses saturate with increasing contrast; second, responses to an effective stimulus are reduced by superimposing a second stimulus (masking); and three, responses suddenly decrease when an initially optimal stimulus is expanded beyond the receptive field of the neuron (size tuning). The mathematical concept they propose simply divides the response of the neuron by a measure of the local contrast from a suppressive field summed with a semi-saturation constant. This formulation is very similar to equation 2. The strength of this study is that after characterizing the model (i.e. determining the correct parameters), they use a different set of stimuli to determine whether their model is capable of predicting contrast saturation, masking, and size

tuning in LGN cells. They demonstrate that their model is indeed capable of doing this, which shows that a model implementing normalization can capture certain response properties of subcortical visual neurons, which form the input of the primary visual cortex (V1). The responses of these neurons show many forms of non-linearities as well, and normalization was initially proposed to account for these non-linearities[3].

Non-linearities in V1 neurons

Neurons in V1 respond optimally to a line with a certain width and orientation. In other words, V1 neurons have spatial frequency and orientation tuning. Within a normalization model, these tuning properties arise from the the linear part of the model, i.e. the way in which a V1 neuron sums its inputs[12]. In mathematical terms, this would mean that the neuron has a specific set of weights (equation 1) that results in responding optimally to a line with a certain width. As V1 neurons obtain their tuning in a linear way, this tuning, together with contrast intensities, should be taken into account when applying normalization to V1 responses. This can be done by weighing the contrast response of the neuron and the contrast response in the normalization pool with a factor that resembles the similarity between the presented stimulus and the preferred feature of the neuron and its normalization pool. As the normalization pool is assumed to be broader tuned than the neuron itself (i.e. it has a wide range of preferred features)[14], this can explain some phenomena of V1 neurons. For example, V1 neurons show cross-orientation suppression[22]. A stimulus that has an orientation orthogonal to the preferred orientation of a V1 neuron does not drive the neuron, but is effective in suppressing the response to the preferred orientation. This can be captured by assuming that the ineffective stimulus only drives the normalization pool, not the neuron itself. When an effective stimulus is presented as well, this drives both the neuron and the normalization pool. The presence of the ineffective stimulus results in a higher activity in the normalization pool than when the effective stimulus would be presented alone, therefore the neuron's activity is suppressed during the presence of both stimuli relative to the presence of only the effective stimulus. The same line of reasoning can be followed when considering the effects of presenting a non-optimal (but still activating) stimulus in isolation, or in combination with an optimal stimulus. In the former case, the stimulus would activate the neuron, but not maximally. In the presence of an optimal stimulus, this stimulus would become suppressive, as it contributes maximally to the normalization pool (some neurons in the normalization pool are assumed to be tuned to this stimulus' feature), but only partially to the activation of the neuron[14].

In sum, by taking into account the tuning properties of both the neuron and its normalization pool, the normalization algorithm can be applied to variety of response properties of V1 neurons. Applying normalization to V1 assumes that its neurons obtain their tuning properties in a linear way, without a normalization step. Nevertheless, tuning properties of MT neurons can be captured by a model that combines several linear summation and normalization steps.

Tuning properties of MT neurons

Simoncelli and Heeger[18] proposed a model to capture the tuning properties of MT neurons. Area MT contains neurons that are sensitive to a combination of speed and direction, also called velocity. This model proposes several steps of linear summation combined with normalization. It assumes that LGN responses are a measure of local contrast as a result of the process of light adaptation discussed above. These responses are summed by V1 neurons (which is captured by equation 1), as discussed above, this summation results in V1 neurons having tuning properties. The response of a V1 neuron is subsequently normalized by dividing its activation by the activation of a normalization pool. The normalized responses of a specific combination of V1 cells are the input of a second class of V1 cells, V1 complex cells. Again, these cells sum their inputs and a set of complex cell's provide input to a MT cell. The MT cell's response is subsequently normalized with respect to the activity of surrounding neurons. This model is capable of reproducing similar response properties as those found in MT velocity selective cells and shows that successive stages of linear summation and non-linear normalization can produce response properties that resemble the properties found at a higher level visual processing areas such as MT.

Discussion

Several applications of the normalization algorithm were discussed briefly. It shows that the same algorithm can be employed to explain non-linearities and response properties at several different stages along the visual processing stream. This strongly suggests that linear summation of inputs and a subsequent normalization captures a process employed throughout the visual system and therefore it may be interesting to consider what its functional benefits might be.

It is likely that the functional benefit of such a process may depend on the exact stage of the visual processing stream at which the normalization occurs. Normalization at the level of the retina (light adaptation) has a very clear benefit of maximizing the sensitivity of photoreceptors by estimating the mean light intensity using responses of neighboring photoreceptors and subtracting the estimated mean from the photoreceptor's response. This positions the steepest part of the photoreceptor's response function at the mean light intensity, thereby making it most sensitive to variations in light intensity around the local mean light intensity. This results in signaling the deviation from this light intensity which is a measure of contrast, as already mentioned above. Interestingly, by discounting the mean light intensity, this measure of contrast is invariant to changes in mean intensity. In other words, the same deviation from the mean intensity is always accompanied by the same neural response, regardless of the absolute value of the intensity. This invariance to a stimulus dimension can be captured by normalization at other stages of the visual processing stream as well. For example, V1 orientation tuning is invariant to contrast[23], and velocity tuned cells in MT are thought to be invariant to spatial pattern (i.e. they respond to motion of whatever object)[14,18]. A final functional benefit could be redundancy reduction. As mentioned above, visual information contains many regularities and is therefore highly correlated across space. Therefore, nearby neurons' responses can serve as a prediction for some neuron's response. The process of subtracting the prediction from a neuron's response in order to obtain an

uncorrelated response could be envisioned as a normalization process. This can be considered as a predictive coding interpretation of these phenomena.

As already mentioned in the introduction, the aim of this overview is not to provide a thorough discussion of the biological mechanism that is captured by normalization. Normalization is an algorithm that captures a property of neural responses and it points towards the activity of nearby neurons as an important part of its biological mechanism, but it does not offer a complete description at the physical level. Nevertheless, it offers an accurate description of some of the properties of the visual system. However, the fact that an algorithm provides a good description of a neural process, does not mean that the biological mechanism implementing this process can be easily linked to the components of the algorithm. This can be appreciated by considering the biological mechanism behind light adaptation. Although normalization captures this process with a very straightforward computation, the biological mechanism is less straightforward. It might be very tempting to assume that light adaptation arises from horizontal cells that laterally connect photoreceptors, but this is not the case. It appears that light adaptation arises from two locations, the first being the photoreceptor itself, the second the relay of information from bipolar to ganglion cells. These sites are mutually exclusive, with the involved location dependent on absolute light intensity[24]. Thus despite having two locations from which light adaptation may arise (i.e. despite being quite complicated on the physical level), it can be captured by a single algorithm (i.e. at an algorithmic level, its properties can be easily captured). Another indication that the biological mechanisms behind normalization are more diverse than the algorithms that capture it seem to suggest is that it is possible that different suppressive phenomena at the same level of visual processing are supported by different biological mechanisms. Furthermore, the exact source of the normalizing activity and the mechanisms through which this normalizing activity modifies neuronal responses is not yet clear for most levels of the visual processing hierarchy[14]. These accounts highlight the simplification that can be achieved by focussing on explaining properties of the visual system in a algorithmic way, rather than a physical way.

Predictive coding algorithms deal with uncertainty and redundancy

Uncertainty of visual perception

Visual perception is inherently ambiguous. Consider that the only visual information available to our visual system is a pattern of activity across our retina's, caused by some unknown object in the external world. Unknown, because our visual system has not yet identified the cause of the sensory information, as this identification (recognition) is part of visual processing. Moreover, each part of the retinal activation can have multiple causes. For example, a very accurate model of a car at a certain distance may cause the same sensory input as a real car at a much larger distance. Thus, retinal activation is inherently ambiguous in the sense that different percepts can lead to the same retinal activation. However, the reverse is true as well: one percept can cause an almost infinite amount of sensory inputs on our retina. First, this can be due to different angles from which we observe the car. As we all probably know, a car looks rather different seen from below than when seen from above, nevertheless we recognize these very different types of sensory inputs as belonging to the

same percept, i.e. a car. Another source of sensory variability are viewing conditions, like illumination of the car. A car seen on a bright sunny day causes a pattern of sensory inputs different from a car seen on a cloudy winter morning. A third immense source of variability is the type of car. A large variety of objects belong to the same category of percepts, i.e. objects ranging from a Volkswagen Beatle to a Rolls Royce all belong to the category of cars. A final source of variability in the sensory signal is the perceptual system itself. For example, this introduces neural noise[5], due to the variability of neural responses, the response of a neuron is never strictly the same despite similar inputs. This means that even if it were possible to view a scene under the exact same conditions for multiple times, the resulting neural signal would still be different. Due to the immense variability in the sensory input associated with each percept, it is impossible for the visual system to employ a one to one mapping of sensory input to the causes of these inputs, i.e. it is unfeasible to store all possible sensory input patterns associated with each percept. Because of these uncertainties associated with perception, it cannot be deterministic, but has to probabilistic. The visual system has to guess which percept causes the sensory data, and can never be entirely sure about the cause of sensory data. This means that there is always some uncertainty associated with what we see. The Bayesian brain hypothesis is proposed to deal with the uncertainty inherent to visual perception. It proposes a set of algorithms that allows the visual system to determine the most likely percept that causes the sensory data[5,6].

The Bayesian brain hypothesis and predictive coding: A solution to uncertainty

When Bayesian inference is applied to perception, it is a method for updating the probability of a percept underlying sensory inputs as additional evidence is gathered. In other words, this means that the brain determines the probability of a percept (cause) given sensory inputs (data) and some prior knowledge about possible causes of this data. The brain determines the most probable cause of sensory data, and perception can be thought of as the process of constructing a probability distribution over a whole range of possible causes of sensory data, assigning a probability to each individual possible cause of sensory data. This probably distribution is known as the posterior distribution. The relation between the posterior probability of a cause, the sensory data and the prior knowledge is given by:

$$P(C|D) = P(D|C)P(C)$$

Where P(C|D) stands for the probability of this cause C given the data D. P(D|C) is the probability of data D given cause C and P(C) is the probability of cause C occurring[5,25,26]. P (C) represents the prior knowledge on causes, and is termed the prior, whereas P(D|C) is termed the likelihood. Let's apply the car example to this equation. Assume someone is walking down a street in a city's center and some object at a distance causes sensory input through his retina. In this case P(D|C) is the probability that this specific pattern of sensory data is caused by a car. Let's assume that it really is a car he is looking at, and it is therefore quite possible that a car causes this sensory data. So, P(D|C) is quite high, let's say 0.95. The fact that this person is walking in a city's center indicates that it is quite likely that he runs into a car at some point, so the possibility of a car 'occurring' is quite high as well, let's say

0.9. In this case, the possibility that his visual computes for the cause 'car' of this current sensory data is 0.95 x 0.9 = 0.855. Combined with the fact that the sum of a probability distribution across all possible causes is always 1, this means that there is not much left for other causes and a car is the cause that has the highest probability assigned to it. Therefore, his visual system correctly 'assumes' it is a car he is looking at. To appreciate how this probability changes with a changing prior or likelihood, consider that same person in the same city center, but now looking at a tree instead of a car. The prior probability of a car would still be 0.9, but the likelihood would be much lower, for example 0.1 (i.e. it is unlikely that a car causes sensory data that resembles that of a tree). The probability of a car as a likely cause for the current percept would be 0.09 in this case. Finally, consider someone hacking his way through a very dense rainforest, and someone managed to get a car somewhere in the middle of this rainforest. The sensory data this person acquires can very well be caused by a car, so the likelihood can be 0.95. But it is very unlikely that someone runs into a car in the middle of a rainforest, so the prior can be 0.2. This results in a probability of a car causing the current sensory data of 0.19.

So, the Bayesian brain hypothesis assumes that prior knowledge (both the prior and the likelihood) combines with sensory data to result in a probability distribution over likely causes of the sensory data. This, however, is not the complete story but just the outcome of an initial 'guess' at the causes of sensory data. This guess comes about by a prior representation of the probability of causes occurring and the probability of some data given the initially assumed cause, but does not have to be necessarily correct or, better said, accurate. Therefore, the initial probability distribution has to be maximized. Predictive coding is an algorithm that can be used to this end. It assumes that the prior and the likelihood (which together form the generative model) generate a prediction of the sensory data, which is tested against the sensory data itself. The difference between the predicted and true sensory data, the prediction error, is used to maximize the posterior, which is in the case of predictive coding similar to minimizing the prediction error. Perception then is determining the cause of sensory data that best predicts the sensory data and which has thus the highest probability of being the cause of the sensory data[6].

Besides finding the most probable cause of sensory data, the generative model used to generate predictions is optimized to increase the efficiency of future perceptual inferences. The prior and the likelihood that constitute the generative model can be conceptualized as probability distributions with a certain shape. This shape determines which statistics (i.e. parameters) are sufficient to describe the probability distribution. For example, in case of a Gaussian shaped distribution, the mean and variance are the parameters that can describe the probability distribution. It are these parameters that are updated[25,26]. Updating the generative model ensures the visual system uses a model that is the best representation of the external environment, given the most recent visual inputs and reflects in that sense a form of learning.

In sum, the goal of the Bayesian brain hypothesis is to obtain a reliable representation of the visual environment, despite the sources of uncertainty associated with visual perception. According to this hypothesis, the brain combines prior knowledge about the external world to determine the most likely cause of sensory data. Predictive coding is an algorithm that offers a possible way in which this can be done. It postulates that predictions

of sensory data are generated and compared with the sensory data. The error of the prediction is used to update the brain's posterior, which results in a new prediction of the sensory data. This updating continues until the predictions of the sensory data match the sensory data to a satisfactory extent, i.e. until the posterior is maximized. Besides maximizing the posterior, the generative model is updated as well in order to process future visual data as efficient as possible. Effectively, within the Bayesian brain hypothesis, the computational goal of predictive coding is to determine the accuracy of a perceptual representation and update this representation accordingly.

Predictive coding and redundancy reduction

Our visual environment contains many regularities. For example, intensities at one position are highly correlated with intensities at neighboring positions. The correlation between intensities at different points is a function of the distance between the two points: the further the two points are apart, the lower the correlations is[20]. Other examples of regularities are the tendency for visual intensities to fall off with increasing spatial frequencies (i.e. most of the visual information is contained in low spatial frequencies)[27] and the tendency of visual intensities to fall off with temporal frequencies (i.e. most visual information is contained in visual patterns that tend to change slowly over time)[28]. Thus, our visual information contains spatial and temporal regularities. Another way of pointing this out is saying that there is redundancy. Effectively, this means that the unique information carried by each individual neuron's response is only a small fraction of its total response, with the remainder of its response reflecting information also represented in spatially and temporally nearby responses. Relaying visual information to higher levels of the visual processing hierarchy without dealing with this redundancy is a very inefficient approach, and information theoretic approaches to perception postulate that the visual system is concerned with improving the efficiency of processing of visual information[7,9,29]. Predictive coding algorithms may offer a potential solution to the problem of redundancy in the visual system.

Besides using a generative model of the visual environment that predicts sensory input, predictions can also be generated using spatial and temporal regularities of visual information. A correlation between two neighboring points in the visual field implies that the neural responses coding for these positions are correlated as well[20]. Exploiting this correlation, the visual system predicts a neuron's response by taking into account neuronal responses from spatially and temporally nearby neurons and subtracts this predicted response from the neuron's response. By subtracting the predictive part from a neuron's response, it is left with only signaling deviations from this prediction, which is the part of the response that cannot be predicted from the surrounding activity and the part of the response that is not correlated with the surrounding responses. By subtracting these predictions from the neural response, the visual system decorrelates the visual signal and removes the redundancy from the visual information. An example of this form of decorrelation is light adaptation (also discussed below). This can be regarded as a process in which the retina exploits the correlation of light intensities to predict neural responses and reduce correlations between them[20]. Furthermore, it has been demonstrated that the retina is concerned with decorrelating its inputs across different spatial scales[27,30], exploiting the fact that spatial frequencies differ

in the amount of information represented. Moreover, the LGN appears to decorrelate inputs across temporal scales[28]. These studies demonstrate that the visual system exploits regularities of visual information to reduce redundancy, as proposed by the information theoretic approach to visual perception. Thus, within an information theoretic approach, the computational goal of predictive coding is mainly to reduce redundancy by exploiting statistical regularities of the visual scene.

Redundancy reduction, however, can also be achieved within the Bayesian brain hypothesis. Here, parts of the visual information that are successfully predicted by a generative model do not have to be represented by units representing visual information, as this information is already represented by units generating the prediction. Therefore, the predictive coding approach assumes that parts of the visual information that can be predicted are subtracted from the visual input, which leaves the visual system with only signaling the prediction errors[29]. Thus, both predicting visual information from statistical regularities or an internal generative model of visual environment can result in redundancy reduction. However, within the Bayesian brain hypothesis, this is not the primary computational goal.

Hierarchical implementation of Bayesian inference and predictive coding algorithms

As already noted above, the Bayesian brain hypothesis deals with the uncertainty inherent to perception. It postulates that perception is finding the most probable cause of sensory data, which is represented by a probability distribution called the posterior distribution. Predictive coding states that this posterior distribution is maximized by comparing predictions of sensory data with the true sensory data. The fact that perception is a probabilistic process means that the uncertainty associated with perception has to be taken into account as well. A growing body of research suggests that human behavior takes into account the uncertainty of sensory input[5,31], which supports the suggestion of perception as being a probabilistic process and the validity of Bayesian algorithms that deal with this uncertainty. As the visual system is a hierarchical system, the Bayesian brain hypothesis and predictive coding have to be considered within a hierarchical context in order to apply their algorithms to visual perception, as is done by Lee and Mumford[26]. Within a hierarchy, the Bayesian brain hypothesis has a generative model at each level of visual processing (e.g. LGN, V1, V2 etc.). The prior at each level represents the probability of a cause given the most probable cause at a higher level. The likelihood represents the possibility of a cause, given the representations at lower levels[6]. Due to this setup each level represents the most likely cause of lower level representations, given these representations and the most probable causes represented at higher levels. Predictive coding again postulates maximization of the posterior distribution by explaining away differences between the predicted and true representations. At the lowest level of the hierarchy, sensory data is predicted. As all stages in the hierarchy are linked together, each level is eventually informed about and shaped by the sensory data[6]. For example, if a corner is predicted at a certain position in the visual field, then this is the outcome of the prior combined with the likelihood, which take into account higher and lower level information respectively. A lower-level prior represents the likelihood of, for example,

lines with a certain orientation, given the predicted corner and the representations at the level below. To maximize predictions, the visual system predicts lower-level representations. In this case, it may predict two lines at a certain angle. This prediction is compared to the true representation of the lower-level. The error between the prediction and true representation is fed forward to update a higher-level posterior and to generate a new, more accurate prediction. In sum, when applied to a hierarchical system, predictive coding assumes that predictions are generated by higher cortical levels that are fed back to lower levels, which subsequently inform the higher levels about the accuracy of the prediction by transmitting a prediction error.

Rao & Ballard[4] have applied this hierarchical interaction of predictions and prediction errors to the processing of natural images. Their model contained artificial neurons at multiple levels. At the lowest level, a natural image was presented, and it was the first goal of the model to predict the image presented (i.e. to determine the cause of the sensory data). The highest level (level 2) of their model generated predictions that were fed into a lower level (level 1), which generated predictions that were compared with the image input (level 0). The error of the predictions was fed into level 1 and level 2 subsequently and was used to improve the predictions. Thus in a model like this, prediction generating neurons are informed about the accuracy of their prediction and they adjust their prediction accordingly. A second goal of this model is to adjust the weighting of their inputs in such a way that subsequent images could be predicted more accurately. The two goals of the model resemble the maximization of the posterior distribution (determine the cause of the sensory data) and the updating of the generative model in order to process future input more efficiently, as discussed above. Furthermore, the neurons at higher levels pool over multiple inputs (i.e. at the first level, neurons receive prediction errors from multiple image pixels, whereas neurons at the second level receive prediction errors from multiple neurons at the first level), which resembles the increasing receptive field size across the visual hierarchy. In addition, a constraint of sparseness was imposed on the network's activity, i.e. the eventual representation of the images should use the least amount of neurons possible[32]. This additional constraint is more related to the principle of efficient coding than it is related to predictive coding and results in artificial neurons having receptive field layouts similar to that of V1 receptive fields.

The application of these predictive coding algorithms to the processing of natural images has interesting results. First, the V1 receptive field profiles obtained strongly suggest that the visual system is concerned with efficient processing, as postulated by information theory. This effectively means that when one wants to process different natural images in the most accurate but nevertheless sparse (i.e. efficient) way, V1 receptive field properties appear to be very good descriptors. Findings more related to predictive coding are those concerning extra-classical receptive field effects. These effects are comparable to the modification of neuronal responses by stimuli that fall outside their receptive field, as discussed under normalization. For example, some of their prediction error units show a response property called end-stopping, which is similar to size tuning discussed above. Here, a neuron responds optimally to a bar when it covers its receptive field only, but when it starts to cover its suppressive surround as well (so it gets longer), the response is inhibited. Rao and Ballard[4] explain this result in terms of predictive coding and suggest that longer bars are more likely

to be part of a natural scene than very short bars. As the generative model used in this study is based on natural inputs, the model predicts that longer bars are more likely to cause sensory data than shorter bars are. In the case of a longer bar being presented, the model correctly predicts its presence and therefore the prediction error is lower. In case of the shorter bar being presented, the prediction of the more natural longer bar is incorrect and the prediction error is higher. Interestingly, disabling the feedback to these neurons eliminated their end-stopping response property indicating that this property is indeed related to the feedback the neuron receives. Another property model neurons exhibited is response modulation due to an orientation contrast between the stimulus at the receptive field center and its surround. Relative to having no stimulus surrounding the receptive field, the presence of an orientation contrast increased the responses of the neuron, whereas the presence of a similar oriented decreased the response. Again this can be explained by natural image properties that have been learned by the generative model. In this case, orientations in one direction are very likely to be accompanied by orientations in the same direction, which is predicted by the generative model when confronted with a certain orientation. These response properties of model neurons have been found in biological V1 neurons as well[4].

Predictive hierarchical models have been applied to other stages of the visual processing stream as well, such as MST, which is a motion sensitive area. Again, when a predictive model is confronted with natural images, artificial neurons start to develop receptive field layouts that resemble those of V1 neurons. Moreover, when presented with visual motion input that resembles MT's input to MST, the artificial neurons start to develop response properties similar to those of MST neurons[33]. Finally, predictive models have been applied to the interaction between LGN and V1 as well and when trained on natural images, they start to resemble the connectivity between LGN and V1 neurons. Moreover, these models can account for the biphasic responses found in LGN neurons in which LGN neurons' optimal stimulus reverses rapidly over time (< 20 ms)[29].

Taken together, these studies demonstrate that when hierarchical predictive coding algorithms are applied to artificial neural networks that process natural images, neurons within the network develop response properties similar to those of biological neurons. This strongly suggests that the brain employs computations similar to those assumed by predictive coding models. Combined with the evidence indicating the importance of uncertainty in human behavior, it also suggests that the brain indeed uses similar algorithms to increase the certainty of its perceptual inferences. Predictive coding models make some specific predictions that can be verified by neuroimaging or neurophysiological studies. Two of these will be discussed here. First, when a stimulus can be predicted by higher areas, responses of lower level prediction error units will be suppressed. Second, as the brain continuously updates its prior expectations to process future information as efficient as possible, consistent changes in the environment should result in changes in the generative models and how the visual system deals with visual information.

Prediction error suppression

Recent neuroimaging studies in humans demonstrate that responses to predictable stimuli are indeed suppressed. For example, when human observers view a random dot motion pattern

(i.e. a patch of dots that may or may not move coherently in one direction) in which the dots are so far apart that only one dot at a time can be present in a V1 receptive field, the coherency of the motion nevertheless modulates the intensity of V1 responses. This means that although V1 neurons are ignorant about the presence and motion direction of other dots in the pattern (as these fall outside the receptive field) the predictability of the trajectory of a dot (due to the coherency of the pattern) nevertheless modulates V1 responses. This is interpreted as suppressing V1 responses through extra-classical receptive field interactions, as they are correctly predicted and therefore generate less prediction errors[34]. Furthermore, the mechanism of prediction error suppression can explain neurophysiological results concerning responses to line orientations that are, depending on the line orientation of their surround, either part of a figure or the background. V1 neurons in macaque almost invariably respond stronger to line orientations that are part of the figure despite being the same stimulus. Within predictive coding, this can be explained as a less successful prediction of the orientations when they are part of the figure, which results in more generated prediction errors. Interestingly, the properties of the response enhancement in the figure condition of this study strongly suggests that feedback projections underlie this enhancement[35]. Support for the suppression of predictable parts of the visual environment has been found in the face-domain as well. A general property of the human sensory systems is that it reduces its response to a repeatedly presented stimulus, called repetition suppression. Here, the relative likelihood of a face presentation being repeated was manipulated and the authors found that repetition suppression reflects the top-down suppression of a prediction error in the context of predictable stimuli, rather than unpredictable ones[36]. Interestingly this latter study highlights not only the importance of the predictability of stimulus, but also that this predictability can be modified by recent experiences the brain has encountered. The tendency for the brain to take into account recent experiences is very clear within the auditory domain[37,38], and highlights the importance of recent experiences to shape the brain's expectations.

Updating the generative model

As already argued above, recent sensory experiences are important in shaping the brain's expectations about the environment as it aims at processing visual information as efficient as possible. So, changing expectations as a function of recent experiences is a logical result of this aim. For example, besides adjusting their sensitivity to the mean contrast on a fast time scale (< 100 ms), retinal ganglion cells also adjust their responses to the full shape of the contrast intensity distribution on a much slower timescale (10 s)[39,40]. Interestingly, it has been suggested that changes in contrast response of retinal ganglion cells are adaptive in the sense that the new receptive field properties of these cells improve predictive coding under the new image statistics[41]. Indeed improving encoding efficiency has been suggested to be one of the functional benefits of visual adaptation[42]. The adaptation of the brain's coding scheme to the statistics of the environment has been found on a larger time scale as well. For example, Berkes and his colleagues[43] related the spontaneous activity pattern of the ferret's brain to the brain's prior expectations and the stimulus evoked activity to the brains causal inferences of sensory data. They found that the spontaneous activity resembled the evoked activity of the

brain increasingly more during development, which indicates that the brain adapts its internal models to natural stimuli even over a period of several months.

Overall, the Bayesian brain hypothesis and predictive coding may provide a comprehensive account of brain function and highlight several important aspects of visual perception: First, it is inherently ambiguous and therefore perception has to be probabilistic. Second, to deal with this ambiguity, the brain constructs and updates its generative model of the environment. Third, the brain utilizes these prior expectations and image regularities to increase the efficiency of visual processing.

Discussion

In conclusion, the Bayesian brain hypothesis and predictive coding provide a general framework of visual processing that can accommodate and explain a variety of empirical findings. However, not all empirical findings may fit the suggestion that top-down processes inhibit neural responses. For example, attention (a top-down) process increases neural responses[44-46]. This contradiction can be reconciled by incorporating attention into a hierarchical framework as a process that serves to increase the certainty of the Bayesian inference, i.e. the certainty of the estimated cause of sensory data[47,48]. Within such a framework, attention acts as a process that counteracts the tendency of predictions to suppress lower level responses, and increases the responses instead. This response increment serves as an increased weighting of sensory data, thereby increasing the impact of attended information on the process that infers the causes of sensory data. Recently, this definition of attention has received empirical support[49], which provides additional support of the validity of the computational goals and algorithms proposed by the Bayesian brain hypothesis and predictive coding. This definition of the intensity of prediction error responses as a weighting of sensory data highlights an important functional implication. It indicates that the parts of the visual scene that are less predictable are treated as more important. Which makes sense if we consider that the unexpected parts of a visual scene are usually more important, for example a fast moving car we did not see coming or the parts of a scene that stand out from the background, such as our stuff lying around on a table.

Although it is not the primary aim of the overview to analyze the Bayesian brain hypothesis at the physical level, it may be interesting to discuss a possible biological mechanism that allows the brain to implement the probability distributions associated with perception. For example, Ma and his colleagues[50] recently demonstrated that population probability codes (PPCs) are very well suited to represent the mean and the variance of a probability distribution. Within these PPCs the probability distribution can be regarded as a Gaussian whose mean is centered on the most likely percept (i.e. neurons whose preferred feature is the most probable one respond most intense) and whose width (variance, i.e. the activity of adjacent neurons) is related to the certainty of the inference. Gain increases (for example, through attention) are in this case related to a narrowing of the Gaussian (i.e. its variance decreases), which can be regarded as an increase in the certainty or precision of the perceptual inference. Thus PPCs rely on the distribution of activity across a population of neurons tuned to certain features. This reliance on neural activity makes the inference sensitive to saturation of neural responses. Therefore, a non-linearity is needed to keep the

neuronal responses within their range of possible response intensities. A non-linearity like divisive normalization as will be discussed below is ideal for this purpose within a PPC[50]. Another way in which normalization may be used within PPC is by normalizing the responses of neurons to a constant, which simplifies extracting the mean and variance from the population code[14]. This suggests that normalization algorithms can supplement those serving the purpose of Bayesian inference, at least at some points along the visual processing stream.

Another possible supplementary role for these models can be appreciated when one considers the way in which V1 receptive field properties come about in hierarchical predictive models. As discussed above, these properties come about when requiring a model to represent natural images as accurate as possible, with a constraint one the amount of neurons necessary to represent each image[32]. This suggests that representing an image as a specific combination of V1 receptive fields is a way in which the resulting neural code is as independent as possible. However, this approach suggests that each natural image is a linear combination of V1 receptive field (also called basis vectors). In other words, a natural image may be represented by a weighted sum of activities of V1 neurons. Here, the summing and weighting is much the same as the linear process captured by equation 1, but concerns V1 neurons that are tuned to a certain spatial frequency and orientation. However, natural images contain many non-linearities as well that result in statistical dependencies, which cannot be accounted for by a linear sum of basis vectors. It has been demonstrated that these statistical dependencies can be removed by an approach similar to normalization[51]. Here, the response of a V1 neuron (as determined by equation 1) is squared and divided by a constant (σ) added to the weighted sum of the squared activity in the normalization pool. The weights assigned to the activity in the normalization pool and the constant are determined using natural image statistics. Interestingly, they demonstrate that adjusting these parameters based on natural image statistics resembles the process of contrast and pattern adaptation, in which neurons adjust their response as a function of persistent changes in the recently encountered visual environment[52].

In sum, predictive coding algorithms capture a variety visual processing properties, thereby providing support for the validity of computational goals they are assumed to serve. Besides that, it provides a functional explanation for different aspects of visual processing such as adaptation, normalization and attention. However, the exact biological mechanisms are not clear and therefore it is possible that different mechanisms at different levels of the visual hierarchy are involved. Nevertheless, recently Bastos and his colleagues[53] proposed that a circuit of neural interactions within and between cortical columns may implement the algorithms suggested by predictive coding, providing an analysis of predictive coding at Marr's physical level.

Conclusion

Normalization and predictive coding are two computational models with rather different computational goals. Normalization aims to explain non-linearities in neuronal responses and does this by dividing the response of a neuron by a measure of activity in the normalization

pool. Predictive coding is an algorithm through which Bayesian inference can maximize its posterior distribution. Besides, it can be a potential solution to the problem of redundancy in visual processing. Both models are well supported by empirical studies and are very likely to be employed by our visual system. This account shows that despite the complexity of the visual system at the physical level, one can capture its behavior in relatively simple algorithms. Moreover, it shows that normalization may capture processes that further elaborate the output of predictive coding algorithms, such as the removal of statistical dependencies[51] and normalizing neural activity within population probability codes[50]. Finally, this account can provide the basis for a discussion on the possibility of normalization and predictive coding implying similar physical mechanisms for some aspects of visual processing.

As is clear from the discussion above, normalization is a model that captures a single property of visual processing: the modification of a neuron's response by neighboring neurons. It does not specify a function of this process, instead it may serve a variety of functions depending on the stage of processing at which it is found[14]. This diversity of function allows it to be applied within a predictive coding framework as well, which also assumes the process of modifying a neuron's response by neighboring neurons in the process of subtracting predictions. The similarity of some of the physical mechanisms captured by normalization and predictive coding is especially clear in the context of light adaptation. This process is very well captured by normalization, but can also be interpreted as the visual system exploiting the statistical regularities of the visual scene and uses nearby light intensities to predict the intensity at a given point[20]. Another indication that predictive coding and normalization may capture similar physical processes is the study by Rao & Ballard[4] discussed above. It shows that neuronal responses change due to different stimuli present in the surround of an artificial cell, which is very similar to the type of response properties captured by normalization. More importantly, Rao & Ballard[4] have suggested that the subtraction of neighboring neuronal activities within their model could have the same result as a divisive normalization procedure, highlighting that normalization and predictive coding may capture very similar physical mechanisms.

It has to be pointed out that an overlap of physical mechanisms captured by normalization and predictive coding is anatomically possible. Although Rao & Ballard[4] seem to emphasize the importance of feedback connections, they explicitly state that their algorithms may capture processes implemented by lateral connections as well. Furthermore, it has been shown that the inhitory surround of V1 neurons mainly reflects the influence of feedback projections from higher visual areas, such as V2, V3 and MT[54]. These feedback connections have been shown to be responsible for the modulation of neuronal responses depending on the stimulus to which they respond being part of a figure or background[35], and may impose a non-linear control on the responses of neurons[54] which is the motivation behind the normalization model[3]. Thus both models can be physically implemented by both feedback and lateral connections.

The possibility that normalization captures the process of prediction subtraction within predictive coding has interesting implications. For example, it implies a potential physical mechanism whose behavior can be captured by normalization. It has been suggested

that cortical layer 2/3 neurons are very well suited to receive feedback predictions and to relay prediction errors to higher levels of visual processing[4,53]. The apical dendrites of these neurons extent into cortical layer 1 where excitatory feedback connections terminate. Layer 1 cells are almost all inhibitory and are likely to be driven by feedback connections and to subsequently inhibit layer 2/3 cells. This inhibition of cortical layer 2/3 cells has been theoretically linked to the process of prediction subtraction[53]. Normalization then, should be able to capture non-linear behavior of cortical layer 2/3 neurons and reflects in this case the inhibitory influence of layer 1 neurons driven by excitatory feedback connections. The positioning of this case of normalization links its physiological mechanism to activity within the gamma frequency range, which is associated with processes in superficial cortical layers[53]. However, cortical inhibition is not the only physical mechanism whose activity manifests as a normalization-like process. Some other physical mechanism can have similar results and potentially underlie different neuronal response properties captured by normalization[14]. Therefore, relating prediction subtraction to normalization suggests that other aspects of cortical processing can be potential implementations of predictive coding algorithms as well.

Furthermore, relating normalization to prediction subtraction implies that the divisive term of the equation should depend strongly on the predictability of the stimulus within the 'receptive field center'. Indeed, Schwartz and Simoncelli[51] proposed a normalization model designed to remove statistical dependencies between V1 neurons. In this model, the divisive term represents a prediction of the neuron's response variance based on the responses of nearby neurons. Moreover, as already noted above, they show that the parameters of their normalization model governing the divisive term can be updated according to the statistics of images presented to their model[52]. This process can be equated to adaptation and within a predictive coding context it may resemble the updating of a prediction generating model, as argued above. In this way, considering prediction subtraction as a normalization process relates adaptation, predictive coding and normalization in a single account of neuronal responses. Furthermore, it suggests that future normalization models may attempt to use the statistics of their stimuli to predict the parameters of their divisive term rather than estimating these parameters through fitting the normalization equation to experimental data.

In sum, relating normalization to subtraction of predictions can direct the search for physical mechanisms underlying both models as an anatomical overlap is presumed. Additionally, it provides hints on the exact nature of normalizing activity as a relation between predictive coding and normalization requires this activity to represent predictions of neuronal activity. Furthermore, this would also mean that normalization could potentially be applied to a range of additional experimental findings that fit nicely within predictive coding, such as repetition suppression, mismatch negativity, and more general processes like figure-background segregation. Besides, every aspect of neural processing that can be captured by normalization could potentially reflect the subtraction of predictions. Future research into the computational models discussed here should take into account their possible overlap and determine the true extent of physical overlap between these models. A true overlap of physical mechanisms captured by these models would enhance their validity. Normalization would have another function ascribed to it, increasing its commonality throughout the brain,

making it more likely to be a canonical computation[14]. For predictive coding, the overlap with a widely employed computation like normalization strengthens its position as an algorithm that captures a neural process playing a pivotal role in visual perception and possibly other modalities as well.

References

1. Van Essen, D. C. in *The Visual Neurosciences* (eds Chalupa, L. M. & Werner, J. S.) 507 - 521 (MIT press, Cambridge, MA, 2003).

2. Marr, D. & Tomaso, P. From understanding computation to understanding neural circuitry. *Neurosciences Res. Prog. Bull.* **15**, 470 - 488 (1977).

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

4. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature* **2**, 79 - 87 (1999).

5. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neuroscience* **27**, 712 - 719 (2004).

6. Friston, K. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* **11**, 127 - 138 (2010).

7. Attneave, F. Some informational aspects of visual perception. *Psychological Review* **61**, 183 - 193 (1954).

8. Atick, J. J. & Redlich, A. N. Towards a Theory of Early Visual Processing. *Neural Computation* **2**, 308 - 320 (1990).

9. Atick, J. J. Could information theory provide an ecological theory of sensory processing? *Network* **3**, 213 - 251 (1992).

10. Robson, J. G. Linear and nonlinear operations in the visual system. *Invesigative Ophthalmology and Visual Science* 117 (1988).

11. Bonds, A. B. Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex. *Visual Neuroscience* **2**, 41 - 45 (1989).

12. Carandini, M., Heeger, D. J. & Movshon, J. A. Nonlinearity and Normalization in Simple Cells of the Macaque Primary Visual Cortex. *The Journal of Neuroscience* **17**, 8621 - 8644 (1997).

13. Carandini, M. et al. Do We Know What the Early Visual System Does? *The Journal of Neuroscience* **25**, 10577 - 10597 (2005).

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

15. Boynton, R. M. & Whitten, D. N. Visual Adaptation in Monkey Cones: Recordings of Late Receptor Potentials. *Science* **170**, 1423 - 1426 (1970).

16. Normann, R. A. & Perlman, I. The effects of background illumination on the photoresponses of red and green cones. *Journal of Physiology* **286**, 491 - 507 (1979).

17. Bonin, V., Mante, V. & Carandini, M. The Suppressive Field of Neurons in Lateral Geniculate Nucleus. *The Journal of Neuroscience* **25**, 10844 - 10856 (2005).

18. Simoncelli, E. P. & Heeger, D. J. A Model of Neuronal Responses in Visual Area MT. *Vision Research* **38**, 743 - 761 (1997).

19. Ohshiro, T., Angelaki, D. E. & DeAngelis, G. C. A normalization model of multisensory integration. *Nature Neuroscience* **14**, 775 - 784 (2011).

20. Srinivasan, M. V., Laughlin, S. B. & Dubs, A. Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London B* **216**, 427 - 459 (1982).

21. Rieke, F. & Rudd, M. E. The Challenges Natural Images Pose for Visual adaptation. *Neuron* **64**, 605 - 616 (2009).

22. Sclar, G. & Freeman, R. D. Orientation Selectivity in the Cat's Striate Cortex is Invariant with Stimulus Contrast. *Experimental Brain Research* **46**, 457 - 461 (1982).

23. Finn, I. M., Priebe, N. J. & Ferster, D. The emergence of contrast-invariant orientation tuning in simple cells of cat visual cortex. *Neuron* **54**, 137 - 152 (2007).

24. Dunn, F. A., Lankheet, M. J. & Rieke, F. Light adaptation in cone vision involves switching between receptor and post-receptor sites. *Nature* **449**, 603 - 606 (2007).

25. Dayan, P., Hinton, G. E., Neal, R. M. & Zemel, R. S. The Helmholtz machine. *Neural Computation* **7**, 889 - 904 (1995).

26. Lee, T. S. & Mumford, D. Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America* **20**, 1434 - 1448 (2003).

27. Field, D. J. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America* **4**, 2379 - 2394 (1987).

28. Dan, Y., Atick, J. J. & Reid, R. C. Efficient Coding of Natural Scenes in the Lateral Geniculate Nucleus: Experimental Test of a Computational Theory. *The Journal of Neuroscience* **16**, 3351 - 3362 (1996).

29. Jehee, J. F. M. & Ballard, D. H. Predictive Feedback Can Account for Biphasic Responses in the Lateral Geniculate Nucleus. *PLoS Comput Biol* **5**, e1000373 (2009).

30. Atick, J. J. & Redlich, A. N. What Does the Retina Know about Natural Scenes? *Neural Computation* **4**, 196 - 210 (1992).

31. Bach, D. R. & Dolan, R. J. Knowing how much you don't know: a neural organization of uncertainty estimates. *Nature Reviews Neuroscience* **13**, 572 - 586 (2012).

32. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607 - 609 (1996).

33. Jehee, J. F. M., Rothkopf, C., Beck, J. M. & Ballard, D. H. Learning receptive fields using predictive feedback. *Journal of Physiology* **100**, 125 - 132 (2006).

34. Harrison, L. M., Stephan, K. E., Rees, G. & Friston, K. J. Extra-classical receptive field effects measured in striate cortex with fMRI. *NeuroImage* **34**, 1199 - 1208 (2007).

35. Lamme, V. A. F. The Neurophysiology of Figure-Ground Segregation in Primary Visual Cortex. *The Journal of Neuroscience* **15**, 1605 - 1615 (1995).

36. Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M.-M. & Egner, T. Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience* **11**, 1004 - 1006 (2008).

37. Garrido, M. I., Kilner, J. M., Kiebel, S. J. & Friston, K. J. Dynamic Causal Modeling of the Response to Frequency Deviants. *Journal of Neurophysiology* **101**, 2620 - 2631 (2009).

38. Wacongne, C. et al. Evidence for a hierarchy of predictions and prediction errors in human cortex. *PNAS* **108**, 20754 - 20759 (2011).

39. Smirnakis, S. M., Berry, M. J., Warland, D. K., Bialek, W. & Meister, M. Adaptation to retinal processing to image contrast and spatial scale. *Nature* **386**, 69 - 73 (1997).

40. Baccus, S. A. & Meister, M. Fast and Slow Contrast Adaptation in Retinal Circuitry. *Neuron* **36**, 909 - 919 (2002).

41. Hosoya, T., Baccus, S. A. & Meister, M. Dynamic predictive coding by the retina. *Nature* **436**, 71 - 77 (2005).

42. Kohn, A. Visual Adaptation: Physiology, Mechanisms, and Functional Benefits. *Journal of Neurophysiology* **97**, 3155 - 3164 (2007).

43. Berkes, P., Orbán, G., Máté, L. & József, F. Hallmarks of an Optimal Internal Model of the Environment. *Science* **331**, 83 - 87 (2011).

44. Martínez-Trujillo, J., C., & Treue, S. Attentional modulation strength in cortical area MT depends on stimulus contrast. *Neuron* **35**, 365 - 370 (2002).

45. Treue, S., & Martínez-Trujillo, J., C.,. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* **399**, 575 - 579 (1999).

46. McAdams, C., J., & Maunsell, J. H. R.,. Effects of Attention on Orientation-Tuning Function of Single Neurons in Macaque Cortical Area V4. *The journal of neuroscience* **19**, 431 - 441 (1999).

47. Rao, R. P. N. Bayesian inference and attentional modulation in the visual cortex. *NeuroReport* **16**, 1843 - 1848 (2005).

48. Friston, K. The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences* **13**, 293 - 301 (2009).

49. Kok, P., Rahnev, D., Jehee, J. F. M., Lau, H., C. & de Lange, F. P. Attention Reverses the Effect of Prediction in Silencing Sensory Signals. *Cerebral Cortex* **22**, 2197 - 2206 (2012).

50. Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nature Neuroscience* **9**, 1432 - 1438 (2006).

51. Schwartz, O. & Simoncelli, E. P. Natural signal statistics and sensory gain control. *Nature* **4**, 819 - 825 (2001).

52. Wainwright, M. J., Schwartz, O. & Simoncelli, E. P. in *Probabilistic Models of the Brain: Perception and Neural function* (eds Rao, R., Olshausen, B. & Lewicki, M.) 203 - 222 (MIT Press, Cambridge, Massachussetts, 2002).

53. Bastos, A. M. et al. Canonical Microcircuit for Predictive Coding. *Neuron* **76**, 695 - 711 (2012).

54. Angelucci, A. & Bullier, J. Reaching beyond the classical receptive field of V1 neurons: horizontal or feedback neurons? *Journal of Physiology-Paris* **97**, 141 - 154 (2003).