The Bayesian-Laplacian Brain

Semir Zeki[1] and Oliver Y. Chén[2]

Laboratory of Neurobiology, University College London and Department of Psychology, Yale University, New Haven, CT, USA

When, to silent sessions devoted to brain thought,
We summon up formulations from endeavours past,
And sigh the lack of many a principle that we sought,
Because those principles were, in our mind, mis-cast,
Lo, for all priors should not be tied in a single Bayesian knot
For biological and artefactual priors each have a separate slot

A (posterior) Bayesian-Laplacian adaptation from Shakespeare's *Sonnets*

Abstract

We discuss here what we feel could be an improvement in future discussions of the brain acting as a Bayesian-Laplacian system, by distinguishing between two classes of priors on which the brain's inferential systems operate. In one category are biological priors (*β priors*) and in the other artefactual ones (*α priors*). We argue that *β priors* are inherited or acquired very rapidly after birth and are much more resistant to varying experiences than α *priors* which, being continuously acquired at various stages throughout post-natal life, are much more accommodating of, and hospitable to, new experiences. Consequently, the posteriors generated from the two sets of priors are likewise different, being more constrained (i.e., precise) for *β* than for *α* priors.

I. Introduction:

We outline below an approach to the Bayesian-Laplacian system as applied to brain studies, which differs somewhat from previous approaches. Our hope is that it may constitute a useful contribution to efforts in neuroscience that address the extent to which the brain uses what may be called Bayesian-Laplacian inferential operations. Our approach is inspired by such knowledge of perception that we have and by past discussions in philosophy, colour vision, and neuroesthetics.

The Bayesian-Laplacian approach is an *evolving* one (Laplace, 1812)(Fienberg, 2006), with physiological and philosophical foundations (Helmholtz, 1867)[3]

[1] Email: s.zeki@ucl.ac.uk
[2] Email: yibing.chen@yale.edu
[3] In the article, Helmholtz described a view which he called the intuition theory (*nativistische Theorie*), that "it is necessary to assume a system of innate

(Bovens & Hartmann, 2005) (Talbott, 2008)(Rosenkrantz, 1977) (Gelman & Shalizi, 2013) and probabilistic, statistical, and computational implications (Good *et al.*, 1966) (Bernardo & Smith, 2008) (Gelman *et al.*, 2004). It summarizes a fundamental inferential principle in which probabilities of occurrence of events are based on *priors* and lead to *posteriors*, which in turn modify inference (Kersten et al., 2004, Knill and Pouget, 2004, Yuille and Kersten, 2006, Clark, 2013) and behaviour (Friston et al., 2011, Botvinick and Toussaint, 2012, Friston et al., 2015).

It is an approach wherein inferential statements, e.g. that the currency of an unstable country will change in value, can be formulated by a simple probability law based upon the current state of that country and historical examples of currency fluctuations with unstable governments. Fundamental to this operation is *belief*, which is intimately linked to priors. The brain must continually update the hypotheses that it entertains about the world in light of the information reaching it and against its current beliefs. Our approach leads us to enquire into different categories of Bayesian-Laplacian priors, the beliefs that they are based on and that they give rise to, and the role that these priors and the beliefs attached to them play in shaping the brain's inferential systems. Our discussion is not exhaustive; rather, we hope that it lays down a basic framework for an alternative approach through which to consider the operations of the brain in a Bayesian-Laplacian context. The major departure in our approach is a distinction between two kinds of priors, Biological (*β priors*) and Artefactual (*α priors*). The former are regulated largely by inherited brain concepts while the latter are subject to acquired (synthetic) brain concepts (Zeki, 2009). This distinction leads us to propose further that the beliefs attached to the two categories of priors must also be distinguished according to category.

As a preamble, it is useful to outline what we believe are some important principles governing the organization of the brain:

1. One of the primordial functions of the brain is the acquisition of knowledge.

2. This knowledge is acquired to take action, but the knowledge comes first.

3. The brain cannot know in advance what conditions or stimuli it will experience or has to acquire knowledge of. It is therefore organized in such a way as to allow it to sample epistemically as many experiences as possible; it

---

apperceptions that are not based on experience, especially with respect to space-raltions. In the same article, he also stated that "the judgment of the sense may be modified by experience and by training derived under various circumstances, and may be adapted to the new conditions." He thus came close to distinguishing between the two sets of priors that we discuss here, although his discussion remains vague and does not give or define either a general or specific framework for distinguishing between priors (a term he did not use) or provide a specific framework of how the "unconscious inference" operates in, for example, colour vision, as we do here.

subjects all inputs to constructive explanation or concepts, which may be regarded as the foundation of the knowledge acquiring system of the brain, because "perceptions without concepts are blind" (Kant, 1787). These concepts are of two kinds, inherited and acquired (Zeki, 2009).

4. Inherited (biological) priors are the result of concepts that we are born with; they are resistant to change even with extensive experience and hence must be distinguished from artefactual concepts.

5. Acquired (artefactual or synthetic concepts) priors depend upon concepts which are formulated postnatally, and which are modified by experience throughout life; they are less constrained than biological concepts.

6. We define the term "concept" separately for biologically inherited concepts and acquired, artefactual, ones. The former may be defined as an inherited abstract "idea" linked to a process that is applied indifferently to incoming signals to generate priors. In colour vision, this would be the biologically inherited concept of ratio-taking, which is applied to incoming chromatic visual signals and results in a colour category which constitutes the posterior and the prior (see Figure 1). It therefore constitutes a generative model which generates sensory consequences (colour) from causes. "Colour [category]" in the words of Edwin Land, "is always a consequence, never a cause" (Land, 1985).

7. An artefactual concept, for example that of a house, conforms more closely to the dictionary definition of "a generic idea generalized from particular instances". Thus the idea of a "house" is abstract and generated from viewing many houses; the concept continues to grow with new experiences. If the brain acquires a generative model of how a significant configuration of stimuli corresponding to actual houses generate the perception of a house, it is in a position to recognize these stimuli as corresponding to its previous experience of houses.

8. Hence, in the Bayesian-Laplacian context, biological priors are much more precise and intransigent than artefactual ones, even though both can lead to almost limitless posteriors. Therefore, the beliefs that are attached to biological priors are more widely shared (i.e., conserved over conspecifics), more independent of culture and learning and also less yielding to experience than artefactual priors.

Historical Note: In 1763, Richard Price published an edited article in *The Philosophical Transactions of the Royal Society* (which at that time had not been, interestingly from the point of view of this article, divided into two sections, representing the biological and physical sciences). The article was by the Reverend Thomas Bayes and entitled *An essay towards solving a Problem in the Doctrine of Chances*. Bayes had died in 1761, and his edited paper was thus published posthumously. It was discovered by the influential French scientist Pierre-Simon Laplace who discussed it in a treatise entitled *Théorie analytique*

des probabilitiés (Laplace, 1812) and formulated the expression now generally associated with Bayes' hypothesis:

$$[H_{\mathcal{C}}|E] \propto [E|H_{\mathcal{C}}][H_{\mathcal{C}}]$$

[See below; the equation in the context of colour vision, which we discuss below, means that $[H_{\mathcal{C}}|E]$, the posterior (percept) of hue H of a colour category $\mathcal{C}$ (where examples of $\mathcal{C}$ can be green, red, and blue), will depend upon the prior which has a belief attached to it, as well as the experiment (experience) conducted, $[E|H_{\mathcal{C}}]$; the posterior thus produced (hue) can then itself act as a prior through further experimentation, $[H_{\mathcal{C}}|E] \rightarrow [H_{\mathcal{C}}]$, resulting in further posteriors which can generate further priors (see also Figures 1 and 2). This is known as Bayesian belief updating and is at the heart of Bayesian-Laplacian evidence accumulation. As an extreme example of a *β prior*, we have

$$[H^0_{\mathcal{C}}|E] \propto [\mathcal{C}]$$

$$[\mathcal{C}|E] \propto [R]$$

where $[H^0_{\mathcal{C}}|E]$, the posterior (percept) of initial hue $H^0_{\mathcal{C}}$, depends only upon the prior $\mathcal{C}$ (colour category), and hence is independent of E (experiment); and $[\mathcal{C}|E]$, the posterior (percept) of colour category, depends only upon $[R]$, the ratio-taking scheme, and is independent of any experiment. Further mathematical proofs are given in the Supplementary Materials.]

By speaking only of the Bayesian hypothesis we, by implication, fail to credit Laplace with the very considerable contribution that he made in establishing the generality of the hypothesis originally formulated by Bayes. We therefore think it right to credit both with this approach, by referring to the Bayesian-Laplacian formulation (Pouget et al., 2013).

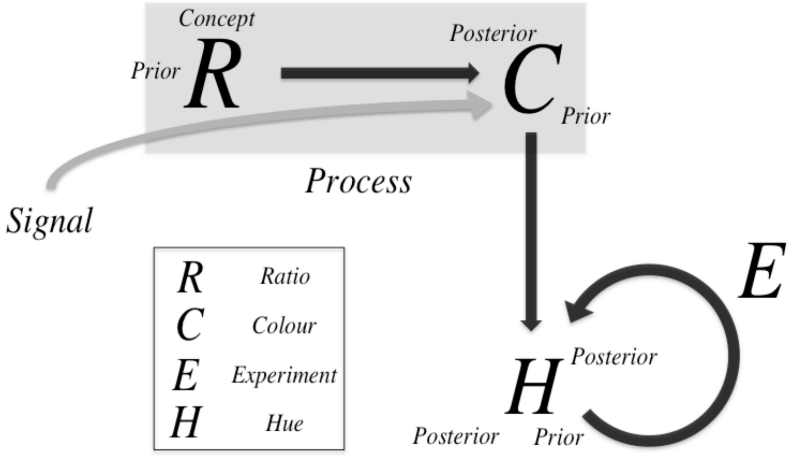


Figure 1: *A schematic representation of the relationship of the ratio-taking process (R) to the generation of the experience of colour (C) as a posterior. The posterior thus generated can act as a prior for generating another posterior (H) through experimentation (E); this latter posterior (H) differs in the shade of colour (hue) from C. Although the ratio taking process R can be thought of as the*

precursor and therefore the prior to C, making C into a posterior, in practice in this case, R is equivalent to C because the process is independent of any experiment (for further details see text).

II. The need for distinguishing biological from non-biological priors

In theory, one could consider all priors under a single category which, subject to experiments or experience, will produce posteriors, as in fact previous discussions of the Bayesian-Laplacian brain have done (Dayan *et al.*, 1995) (Rao & Ballard, 1999) (Lee & Mumford, 2003) (Kersten *et al.*, 2004)(Knill & Pouget, 2004) (Yuille & Kersten, 2006) (Friston *et al.*, 2011) (Botvinick & Toussaint, 2012) (Clark, 2013) (Pouget *et al.*, 2013) (Friston *et al.*, 2015). In practice, the distinction between the two sets of priors is significant enough for experience to operate on the two categories of priors in different ways; the scope of experience to modify is more limited for *β priors* and the beliefs attached to them than for α *priors*; consequently, the beliefs attached to the *β priors,* and the inferences drawn from them*,* are also more biologically constrained than the ones attached to *α priors*.

The classification of priors into two broad categories proposed here is based in part on the Kantian system and in part upon our modification of it (Zeki, 2009). Kant wrote in *The Critique of Pure Reason* (Kant, 1781) that, "perceptions without concepts are blind", arguing that all inputs into the mind (in our case the brain) must be somehow organized by being interfaced through concepts. In *The Critique of Judgment,* he nevertheless proposed that some sensory inputs are not interfaced through concepts; among these were signals from objects that could be categorized as beautiful, as opposed to those which had a utilitarian value. The former were "purposeful without a purpose" and the perceiver usually supposed (believed) the operation of a universal belief through which what s/he had perceived to be beautiful would also be perceived to be beautiful by others. We differ from this classification by supposing that *all* percepts, even those pertaining to beauty, are interfaced through concepts but we distinguish percepts that are grounded in inherited *β priors* from the post-natally acquired *a priors* , which are acquired postnatally (Zeki, 2009). There is good reason to suppose that the inherited priors, which make (Bayesian and Laplacian) sense of the sensory inputs into our brains are much more similar between humans and also far less dependent upon culture and learning than the acquired ones . Hence one cardinal distinction between the two sets of priors is that an individual can reasonably suppose that a biological prior, such as colour, is an experience that s/he shares with the great majority of other individuals, regardless of race or culture, a characteristic not shared by the artefactual priors (e.g. appreciation of sushi, or the beauty of a temple or a cathedral).

This distinction is not commonly made and there is therefore no general agreement as to what disambiguates biological and artefactual priors. There are probably other attributes that fall into the biological category, for example that of motion, but we do not discuss these extensively. Rather, we give examples of what most would agree fall into different categories – colour and faces on the one hand, and the many artefacts such as buildings or cars on the other. Other

beliefs, quite distinct from objects, also fall into the artefactual category; for example, the supposition that a government with a hardline policy on health may, if elected, lead to a rapid change in the value of companies producing medicines. Much of the distinction that we make is based on common human experience. Therefore, although there may not be common agreement on the ontology of biological and artifactual priors, their phenomenology speaks to a clear qualitative distinction. We are effectively claiming that biological priors – that are conserved over generations and cultures – have a greater precision than the more accommodating (empirical, artefactual) priors we call on to assimilate experience that is unique to our time and place.

Our hope is that the differentiation we thus make will be a stimulus for further discussion on how the brain handles these two different categories.

We address the distinction between biological and artifactual priors in terms of visual perception, about which relatively more is known and with which we are better acquainted. We give two examples, among many, of *β priors*, those belonging to colour and face perception.

III. Colour vision:

III A: Colour as a β prior generated from an inherited brain ratio-taking concept

Colour represents perhaps the most extreme form of a *β prior* which is the result of an inherited prior concept, that of ratio-taking, which leads to constant colour categories (see below). Colour is an experience, which we refer to as its *β prior* and from which hues may be generated which, though belonging to the same colour category, differ in appearance (in shade of colour). These hues become posteriors and act also as new priors from which further hues (posteriors) can be generated through experiments and experience. Colour categories and hues are also biological signalling mechanisms allowing the rapid identification of objects, biological or otherwise, by one of their characteristics. If the colour of an object or surface were to change with every change in the illuminant in which it is viewed, then colour would no longer be a useful biological signalling mechanism, because the object can no longer be identified by its colour alone. To make of colours an autonomous identifying mechanism, *they must be stabilized and be immune, as far as possible, from the de-stabilizing effect of a change when the wavelength composition of the illuminant in which objects and surfaces are viewed changes.* Physiologically, the brain undertakes a ratio-taking operation to render colours constant, and hence stabilize the world. Evidence shows that there are specific brain pathways and a specific visual area, area V4 and the associated V4a, that are crucial for colour perception(Zeki, 1973)(Zeki, Watson, & Lueck, 1991)(Bartels & Zeki, 2000), damage to which leads to the syndrome of cerebral achromatopsia (Meadows, 1974)(Zeki, 1990). It is important to emphasize that V4 not only responds specifically to categories of colour but also to hues, or shades of colour (Zeki, 1980)(Stoughton & Conway, 2008) (Brouwer & Heeger, 2013) and that the representation of colour within V4 can be independent of form (Zeki, 1983)(Lafer-Sousa, R, Conway, BR, Kanwisheer,

2016). It is therefore likely that area V4 is pivotal to these operations (Bartels & Zeki, 2000), which is not to say that it acts in isolation; it does so in co-operation with the areas it receives signals from and projects to, together with the reciprocal connections between these areas. Colour is thus an inherited prior which is severely constrained even though, paradoxically, it also allows an almost unlimited variety of chromatic experiences and therefore posteriors (in terms of hue - see below).

***III B. The brain's ratio-taking system for generating constant colour categories*:**

Colour is a brain construct (Zeki, 1984), which is on the one hand a prior for generating hues and on the other a posterior belief, the product of an operation based on an inherited brain concept, that of ratio-taking; that operation compares the wavelength composition of light reflected from one surface with that reflected from surrounding surfaces, thus providing a ratio for light of every waveband reflected from a viewed surface and from its surrounds(Land, 1974) (Land, 1986) (Land & McCann, 1971)(Zeki, 1984). Note that we can refer to colour as a posterior, i.e. the product of an operation, that involves ratio-taking. However, we have shown that, using the argument presented here and the mathematical proof given in Equation (1) of the supplementary materials, this process is independent from any experiment; namely, a colour category is deterministic of (only dependent upon) the ratio-taking operation. Therefore, in essence, the colour category and the ratio-taking operation are equivalent. Hence within a Bayesian-Laplacian framework, the ratio-taking operation and the colour category constitute the prior and posterior, respectively (see Figure 1).

The ratios thus produced never change (see Figure 1) – they constitute fundamental invariants in a world composed of colourful objects. It is useful to discuss briefly here why this should be so in the context of what the constants in nature are, with respect to how these constants generate constant colour categories.

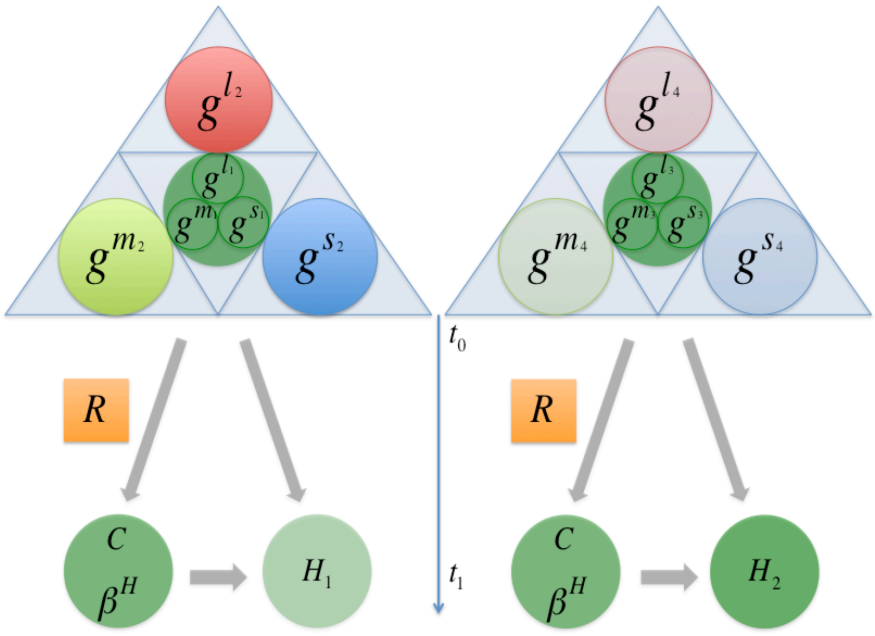


Figure 2: *The Bayesian brain procedures for colour vision. R refers to the ratio taking scheme which leads to our perception of the colour category C, which acts as the biological prior for hue (H). When the wavelength composition of the illuminant changes (as shown to the right), the colour category C remains the same but the hue changes to* H_2. *Note that that the illuminant changes at* t_0, *leading to the same colour category C and the new hue* H_2 *at* t_1. *For details, see text and for proof of the equations see supplementary material.*

The unvarying property of surfaces in terms of colour vision is their reflectance, namely the amount of light of any given waveband – in percentage terms – that a surface reflects in relation to the light incident on it. For a given surface, this percentage never changes. Hence, one finds that the ratio of light of any waveband reflected from a given surface and from its surrounds also never changes, regardless of the variation in the amount of light reflected from an object. If the intensity of light of any given waveband reflected from a surface is increased, the intensity of light of the same waveband coming from the surrounds also increases, and the ratio thus remains the same. By extension, the ratios of light of all wavebands reflected from a surface and from its surrounds also never change.

Take as an example a green surface which forms part of multicoloured (natural) scene as in Land's colour Mondrians, and thus surrounded with many patches of other colours, with no patch being surrounded by another patch of a single colour (Figure 1 shows a much simplified version). Let us suppose that the green patch (g) reflects x per cent of the long-wave (red) light, *l*, incident on it, y percent of the middle wave (green) light, *m*, (green) and z percent of the short-wave (blue) light, *s*, incident on it. The surrounds, having a higher efficiency for reflecting long-wave light, will always reflect more and there will be a constant ratio in the amount of red light reflected from the green surface and from its surrounds. Let us call this ratio g^lThe surrounds will have a lower efficiency for reflecting green light and hence there will be another ratio for the amount of green light (g^m) reflected from it and from the surrounds, and a third ratio for the amount of blue light g^s [4]. When the same natural scene is viewed in light of a different wavelength composition, the amount of light of different wavelengths reflected from a surface and from its surrounds will change, often significantly (as, for example, when a scene is viewed successively in tungsten light, in fluorescent light or in sunlight) but the ratios in the amount of light of different wavebands reflected from the centre and from the surrounds remain the same. Formally, let us consider the green patch viewed under two different illuminants, where the first one has g^{l_1}, g^{m_1}, and g^{s_1} amounts of long, middle, and short wave light reflected from the green patch, and the second one has g^{l_3}, g^{m_3}, and

[4] For brevity, we restrict ourselves to long, middle and short-wave light, without giving the peak values along the spectrum; in practice, and under natural viewing conditions, a surface will reflect light of many wavelengths, but there will be a (constant) ratio for light of any wavelength reflected from the centre and surrounds.

g^{s_3} reflected from it. The first illuminant results in (g^{l_2}, g^{m_2}, and g^{s_2}) and the second in (g^{l_4}, g^{m_4}, and g^{s_4}) ratios of long-, middle-, and short- wave light reflected from the green patch and from its surrounds, respectively. Then, mathematically, we have:

$$Ratio_{long} \coloneqq \frac{g^{l_1}}{g^{l_2}} \equiv x \equiv \frac{g^{l_3}}{g^{l_4}};$$

$$Ratio_{medium} \coloneqq \frac{g^{m_1}}{g^{m_2}} \equiv y \equiv \frac{g^{m_3}}{g^{m_4}};$$

$$Ratio_{short} \coloneqq \frac{g^{s_1}}{g^{s_2}} \equiv z \equiv \frac{g^{s_3}}{g^{s_4}}$$

where the ratios x, y, and z remain the same, regardless of the precise wavelength-energy composition of the light reflected from the green patch and from its surrounds.

It is through such a that the brain builds *constant colour categories.* These constant colours constitute the *β priors*; namely, the prior beliefs that sensations are caused by coloured objects in the particular (ratio preserving) fashion described above. This belief is a classic example of Kant's statement in his *Prologemena* (Kant, 1783) that "The Mind [brain] does not derive its laws (*a priori*) from nature but prescribes them to her".

The brain likely uses such a ratio taking-system, based on the above constants, to construct constant colour categories, which constitute both the experience and posterior beliefs (the two are here interchangeable, see above). There is no physical law that dictates that such ratios should be taken; it is instead an inherited brain law and, given the widespread use of colour constancy in the animal kingdom, we make the assumption that similar mechanisms, or very nearly so, are used in species as far apart as the goldfish (Ingle, 1985) and the human (Land, 1974).

No amount of visual experience can modify the colour categories (*β priors)*; in fact they cannot even be modified by higher cognitive knowledge. For example, green leaves reflect more green than red light at noon and more red light than green at dawn and at dusk; but they are always perceived as green, even in the face of knowledge that they are reflecting more red light under certain conditions. Hence the *β priors* in colour are extremely stable and un-modifiable with experience. Technically, they are endowed with precision. Priors exert a more constrained effect over posteriors when they are more precise, as in biological priors. We use precision here in a technical fashion to denote the confidence afforded to priors; precision is the inverse dispersion or uncertainty encoded by a probability distribution (e.g., the variance of a Gaussian distribution). We define precision as 1/variance. In normal colour vision, given the *B prior*, the variance is very close to 0 and the precision therefore approaches an infinite value.

III C. The experience (experiment) with (constant) colours

This does not mean to say that the shade or hue of the green patch remains constant. The *hue* will change (a) with every change in wavelength composition and (b) can be much influenced by the surrounds. Thus, if the green surface is reflecting more middle wave light than long and short wave light, it will appear a lighter green than when it is reflecting more long-wave than middle or short wave light, when it will appear a darker green (Figure 1, right). Similarly, through the process of colour induction, the saturation of the green surface can be enhanced by surrounding it with red patches exclusively. In brief, the constant colour category (green) constitutes the primal experience of colour (β *prior*), while the hue constitutes the posterior derived from it; the direction in which the posterior (hue) changes can be predicted with high accuracy through experience (see below) and it is indeed this knowledge gained through experience and experimentation that artists use constantly. It is interesting to note that the cortical response to colour in pre-linguistic infants (5-7 months) measured by near infra-red spectroscopy indicates that there is a significant increase in activity in occipito-temporal regions (presumably including area V4) with between-category (colour) alterations but not with within-category (hue) alterations (Yang *et al.*, 2016); similar results have been reported in other studies comparing infants and monkeys (Bornstein *et al.*, 1976), consistent with the view expressed here, that constant colour categories are the priors and that hues are the posteriors.

Another posterior is provided by attaching colours to definite shapes or conditions. Let us take green as a β *prior*. In our culture, green is linguistically attached to a number of objects and these, through experience, become posteriors. Thus leaves are normally green, but there could be conditions in which they are not. When the Fauvists wanted, in their words, to "liberate" colour to give it its maximal emotional intensity, they simply attached unusual colours to common objects such as trees or rivers. Here the new and unaccustomed colour-form combination becomes the posterior, with the artist confident in the knowledge that the β *priors* can be used to generate such posteriors. This usage does not affect the β *priors*; it simply changes the way in which they are used to modify the posterior.

We propose the following *Bayesian-Laplacian Brain Theorem* (see Appendix for mathematical details), which summarizes the above example theoretically: the β *prior* (e.g. the constant colour category generated from the ratio-taking operation detailed above) generates a *posterior* of any quantity of interest (for example, hue in the above instance). Let us refer to this as H_G^0, where the subscript G refers to the green colour category, and the superscript 0 indicates that it has an initial hue. The experiment conducted (for example, adjusting the illuminant in a colour experiment or surrounding the green patch with a patch of a single colour, say red) will lead to a posterior hue which, though still belonging to the colour category of green, will differ in its shade of green (and hence hue) from H_G^0. We will refer to this as H_G^1. As one experiences different shades of green (different hues) when one view the same scene in different illuminants or, in the example of the green patch above, changes its surrounds, and notes the

nature of changes in the illuminant and/or the surrounds of the green patch being studied, so any number of different shades of green can be generated and experienced. These posteriors (H_G^i, where i = 0, 1, 2, …, can be updated continuously and iteratively in one's life; knowing more about the illuminant or the spatial configuration of a stimulus, one can therefore make inferences with a high degree of accuracy and reliability.

III D. The 'belief' with respect to colours

A definition of 'belief' might be adapted from its ordinary dictionary definition, namely "a feeling of being sure that someone or something exists or that something is true" (*Webster's Dictionary*), or "confidence in the truth or existence of something not immediately susceptible to rigorous proof (*Dictionary.com*) or that "the experience (of colour) will always be true", even when we are not remotely aware of the operations that lead to the experience. The belief with respect to colours is subtle; it consists of unconscious knowledge and can be illustrated with respect to the green patch referred to above. *Helmholtz introduced the concept of the "unconscious inference" to account for this but he qualified it in a way from which we depart, for he supposed that judgment and learning enter into the "unconscious inference". We believe that the inference is due to an automatic, inherited application of a concept, that of ratio-taking (but also see above).* A viewer 'knows unconsciously' that the green patch will look green if it reflects more green light than its surrounds, regardless of the actual amount of green light reflected from it and regardless of whether s/he is acquainted with the object or had never seen it before, thus precluding judgment and learning. Anyone armed with such knowledge can predict the colour of a surface, even before seeing it. Thus, a red surface is one that will reflect more red light than its surrounds and a blue surface is one that will reflect more blue light than its surrounds. A white surface will reflect more light of all wavebands than its surrounds while a black one will reflect less light of all wavebands than its surrounds, regardless of the actual amount of light reflected from it. This belief system is quite rigid and not easy to manipulate. Moreover, any perceiver can make the reasonable assumption that the colour perceived under any given conditions will be the same for all, regardless of culture and upbringing.

This can be readily established by matching the colours perceived with Munsell chips. Hence the belief is universal. Because of this assumed universally shared belief, Kant would no doubt have referred to this as an experience which is not interfaced through a (prior) concept. But, as we have shown, it is in this instance an experience generated through an, the ratio-taking operation, executed by the brain to stabilize the world in terms of colour categories. We leave out of account here the vexed and unsolved problem of qualia, of whether the quality of green that one person perceives is identical to that perceived by another.

It is also worth pointing out that knowledge that a green leaf reflects more red light (as is common at dawn and dusk) will not, and cannot, modify one's experience of its colour as green, provided that the leaf is being viewed in a natural context (thus allowing the brain's ratio-taking system to operate), again calling into question Helmholtz's supposition that judgment and learning are

critical for determining the colour category. In this sense, the experience and the belief attached to it are biologically constrained.

IV. Faces – a category of B priors

It is generally agreed that there are special areas of the brain that are necessary for the perception of faces, including an area located in the fusiform gyrus known as the fusiform face area (FFA) (Sergent et al., 1992)(Kanwisher et al., 1997) damage to which leads to the syndrome of prosopagnosia. We note that the FFA is active when faces are viewed from different angles, hence implying a certain degree of *face constancy* (Pourtois *et al.*, 2005). Another area critical for faces is located in the inferior occipital cortex and known as the occipital face area (OFA) (Peelen & Downing, 2007)(Pitcher, 2014) while a third area, located in the superior temporal sulcus, appears to be important for the recognition of changing facial expressions (Haxby et al., 2000). These may not be the only areas that are important for face perception. It has been argued that the recognition of faces engages a much more widely distributed system (Ishai *et al.*, 2005); it has also been argued that cells responsive to common objects, in addition to faces, can be found in an area such as FFA. Whatever the merits of these contrasting views, they do not much affect our argument, given the heightened susceptibility of faces to distortion and inversion and the relative resistance of objects to similar treatment (Zeki & Ishizu, 2013, for a review); this would argue in favour of our general supposition that *β priors* must be separated from *α priors*, whether the representation of objects and faces occurs in the same or in different brain areas (for a general review, Zeki & Ishizu, 2013).

It is generally also agreed that the capacity to recognize a certain "significant configuration" (Zeki, 2013) as constituting a face is either inherited or very rapidly acquired, within hours after birth (Goren *et al.*, 1975)(Johnson *et al.*, 1991), although there has been much discussion as to what it is in the configuration that is instantly recognizable (see discussion in (Zeki & Ishizu, 2013). It is significant that this preference of the new-born for looking at faces is not found when line drawings of real faces are used (Bushnell *et al.*, 1989) emphasizing the pre-eminence of the *biological* concept of face.

Any departures, even minor ones, from the significant configuration that constitutes the (possibly) biologically inherited and accepted concept of a face is rejected and never incorporated into the concept of a normal face. The cortical response to faces is itself very exigent in terms of the significant configuration that it will respond to optimally; mis-aligning the two halves of an upright face delays and increases the typical N170 negative deflection obtained following facial stimulation, but this delay and increase are not quite as strong for inverted faces(Ishizu *et al.*, 2008), which are immediately classified as having an abnormal configuration, and thus not belonging to the biological category of faces. It is, we think, very difficult to produce a biologically viable *posterior*, and a belief attached to it, from an inverted face (unlike buildings – or other artefacts –

see below) and, even if produced, is unlikely to be durable. The same is true of the expression on a face, with certain expressions being immediately recognizable as comforting or loving and others leading to different emotional apprehensions. The *posterior* that results from this *β prior* through experience is thus similarly circumscribed, since any departure (as produced by inversions, for example) would mean that the brain will either not classify it as a normal face, or that it will only be temporarily classified as a face, or that it will be classified as an abnormal face, without leaving a permanent posterior.

Any posterior generated from a face *β prior* must therefore be strictly linked to what is a normal significant configuration which constitutes a face. A child, for example, begins to learn to associate certain expressions on a normal face with certain social interactions – whether, for example, someone is enjoying one's company or is bored by it, whether small inflexions represent doubts or threats, and so on. But it is unlikely to associate the expressions on an inverted face with a permanent posterior, since an inverted face disobeys the inherited brain concept of the significant configuration that constitutes a face. This does not mean that posteriors related to faces cannot be of an unusual nature – for example a continual smile linked to wicked behaviour on the part of an individual may lead the perceiver to establish a different posterior from the same *β priors* for that individual. But here again, a perceiver is unlikely to form a permanent posterior of a smile linked to wickedness if the face is mis-aligned or inverted.

It is interesting to note that, in his effort to give what he called "a visual shock", the English painter Francis Bacon subverted the brain's *β priors* in terms of faces, and took to painting highly deformed faces, which depart significantly from the significant configurations that constitute the normal *β priors* for faces (Zeki & Ishizu, 2013); however prolonged the viewing, these never become accepted as normal faces. Indeed, the viewing of stimuli in which inherited concepts of face (and space) are deformed and violated leads to significant activation in frontoparietal cortex, whereas the viewing of "deformed" or unusual configurations of common objects such as cars do not. Even daily exposure to deformed faces and deformed objects for 1 month does not lead to a significant change in activation patterns for both categories, suggesting that such biological concepts are stable at the neural level, at least within a time frame of 1 month (Chen & Zeki, 2011). This neurobiological demonstration is consistent with our proposed subdivisions of priors into the biological and artefactual categories.

IV A. The 'belief' attached to faces:

We now outline in general terms the biologically based initial belief attached to normal faces. It is constrained by the fact that a face must contain a certain number of features such as eyes, nose, mouth etc, set out within certain proportions and symmetrical relations to each other which, together, constitute a significant configuration typical of a face. An absence of any of these features or any significant violation of these proportions or relations will automatically depart from such a belief, and lead to its classification as abnormal. There are of course many ways in which faces can be represented; they can, for example, be

represented in terms of straight lines in a drawing. But such, though recognized as representing a face, will be immediately classified as a drawing and therefore not a biological face. This shows how constrained such a belief and the *β prior* attached to it is. In terms of generality, one person's belief that the object s/he is seeing is a face and that all others will also perceive a face in that configuration is a sound one and makes that belief general. Just like colour, it therefore has universal validity.

V. Artefactual (a) priors

By artefactual priors, we refer to the many constructs – from houses and cars to ordinary utensils and tools – for which there is no inherited brain concept. Instead, the brain acquires a concept of these objects through experience and consequent updating of empirical priors; these are continually modified throughout life. These empirical priors are also strongly culture dependent. In medieval times, people had no concept of a car or a plane. Since their introduction, there have been many modifications of these constructs, and the concepts attached to them have changed accordingly. The concept of a plane that someone living in the 1930s had, for example, did not include jumbo jets equipped with jet engines; these have been added to the overall concept of a plane since. There are, of course many other examples one could give, including the use of knives and forks and chop-sticks, which differ between cultures and times. The formation of such concepts is strongly dependent on experience and culture, which distinguishes them from biologically inherited concepts (see Figure 2).

Crucially, acquired or empirical priors emerge de novo and are driven by experiences that are unique to any individual in any given lifetime, although there may be, and usually are, population level similarities. They therefore are necessarily less precise and more accommodating than biological priors. This follows because they are designed to be modified by experience.

It is now generally accepted that there is a complex of areas, known as the lateral occipital complex (LOC) which is critical for object recognition (Grill-Spector, 2003). Even though it has been argued that the so-called face areas may not be as specific to faces as originally supposed (see above), and that cells in them may encode objects as well, including ones which we would classify under artefactual categories, the differential response to faces and objects when inverted suggests that they are processed differently. Moreover, neural sensitivity to faces increases with age in face-selective but not object-selective areas of the brain, and the perceptual discriminability of faces correlates with neural sensitivity to face identity in face selective regions, whereas it does not correlate with a heightened amplitude in either face or object selective areas (Natu *et al.*, 2016). There is no definitive evidence about when infants begin to recognize objects or whether they recognize faces before recognizing objects. Indeed, it has been shown that infants can recognize differences between shapes even at 1 month where the outside contour/shape is static and identical, but where the inside smaller shapes are different to each other in each image if, significantly, one of the smaller inner shapes is jiggled or moved (Bushnell *et al.*, 1989); this may, in

fact, introduce a biological *prior*, that of motion, into the recognition or inference process.

Such results, together with common experience, justify a neurobiological separation between the two categories, faces belonging to the biological category and objects to the artefactual.

VI. A biological prior that makes artefactual priors possible?

While the emphasis in this article is on separating biological from artefactual *priors*, it is worth asking whether, at the earliest recorded stages after birth, one can postulate the presence of a general biological *prior* that leads to artefactual priors, which then assume an autonomy of their own. The common view is that there is one category of cell in the visual brain, the orientation selective (OS) cell, discovered by Hubel & Wiesel (1962), which is the physiological 'building block' of all forms. This is a plausible argument entertained by both physiologists as well as artists like Piet Mondrian (see Zeki, 1999). Evidence from physiological and clinical studies of form perception studies shows that, while the OS cells and hence the machinery for constructing forms must be present at birth, it requires nourishment in the early stages after birth to mature. The most comprehensive studies come from the work of Wiesel and Hubel, who showed that OS cells are present at birth but that depriving the animal (cat or monkey) at a critical period after birth blights their visual capacities for considerable periods, perhaps even permanently, thereafter (Hubel & Wiesel, 1977). Observations in humans deprived of vision at birth through congenital cataracts, with vision restored later in life after successful operations, confirms that visual nourishment during an early ' critical' period is necessary for a normal visual life (for a review, see (Zeki, 1993).

By contrast, a normally nourished visual brain can subsequently recognize and categorize many different shapes, even those that have not been seen before. Hence, one could consider that OS cells are the given biological *priors*. In accepting the common supposition that OS cells constitute the physiological 'building blocks' from which all categories of objects (including faces) are constructed, one must nevertheless acknowledge that (a) orientation selective cells are widely distributed in different, specialized, visual areas of the brain (Zeki, 1978), and that the OS cells of V1 may not be the sole source for the neural construction of objects, especially since OS cells in visual areas outside V1 survive the destruction of V1 (Schmid *et al.*, 2009), thus showing that their properties are not wholly dependent upon input from V1. Thus, OS cells in different visual areas may contribute to form construction in different ways. Moreover, unlike what is commonly posited, V1 is not the sole source of the 'feed-forward" visual input to the rest of the visual brain; the specialized visual areas, including areas with high concentration of OS cells as well as areas specialized for face and object perception, receive two further "feed-forward" inputs, from the lateral geniculate nucleus and the pulvinar (Zeki, 2016)) and are activated with the same latencies, post stimulus presentation, as V1 (Shigihara & Zeki, 2013)(Shigihara & Zeki, 2014a)(Shigihara & Zeki, 2014b). Hence a strictly hierarchical organization for form (as is commonly supposed), in which cells within the brain's form system acquire increasingly more complex properties,

enabling them to respond to complex objects and faces is probably unlikely. Rather, there appears to be multiple hierarchical systems, which operate in parallel and which are task and stimulus dependent (Zeki, 2016).

There is another difficulty in considering OS as being the universal biological *priors*. One cannot build a definitive posterior from a single or from multiple oriented lines. If faced with either, what would the posterior be? This is quite unlike colour, where certain ratios of wavelength composition of light reflected from a patch and from its surrounds determines, *ineluctably*, a certain *prior* in the form of a certain constant colour category, from which posteriors, in the form of hues, can be elaborated. We are, we believe, therefore justified in supposing that orientation selectivity cannot be a biological *prior* for all forms, as most suppose. Rather, OS cells in different areas may be used to construct different forms or different categories of form, which then act as distinct priors from which posteriors can be generated. Indeed a line need not be a means toward a more complex form; it can exist on its own, as artists have so frequently demonstrated. Moreover, there is no belief that can be attached to single oriented lines, except in the narrow sense that they can constitute, either singly or in arbitrary combination, forms in themselves, as Alexander Rodchenko (1921) argued when he wrote "I introduced and proclaimed the line as an element of construction and as an independent form in painting". He added, "…the line can be expressed in its own right, as the design of a hypothetical construction [and can have] a status independent of what is actually taking place, and becomes an abstraction" (Rodchenko, 1921) (our ellipsis). Many artists since then have emphasized the primacy of the line in their work.

Hence, there is no universal belief that is attached to how single oriented lines can be combined. There is also no universal belief attached to what significant configuration constitutes a given category of object. The configuration of houses, as places of habitation, differs widely in different cultures – from igloos to huts to skyscrapers and even to inverted pyramidal buildings (Figure 4). One cannot make the assumption that huts are the universal mode of habitation or that inverted buildings depart from the concept of habitation. Rather, the latter are absorbed into the concept (i.e. generative models) of habitation through experience.

Conclusion:

We have here given a general account of what we believe is an important distinction to be made when considering the brain as a Bayesian-Laplacian system. For simplicity, we have concentrated on extreme examples, ones which we have better knowledge of; namely, that of colours and faces for the *β priors* and of common artefacts for the *α priors*. This naturally leaves out of account a vast territory in which both priors may be involved. Laplace himself delved into questions of average mortality and the average duration of marriages. The list can be extended to include social interactions as well as economic activity in which the unfortunately un-studied *β prior* of greed may play a crucial role, in addition to *α priors*. An example of the latter, which plays a role in economic calculations, is the recognition of political decisions that influence monetary

values, which would fall into the artefactual, α, category. In these, and many other human activities that involve making inferences based on a set of beliefs, the distinction between the two categories of priors is, we believe, important. Finally, the distinction between biological and artefactual priors can also be extended to aesthetics, since aesthetics pertaining to biological entities such as faces or bodies, are similarly constrained by the configurations that constitute them (Zeki 2009)(Zeki, 2013).

We have restricted ourselves here largely to the visual brain, but hope to deal with other brain processes that are subject to Bayesian–Laplacian operations in future papers.

Acknowledgment:

We are very grateful to Karl Friston, Will Penny, and Stewart Shipp for their insightful comments on an earlier version of this paper.

References:

Bartels, A. & Zeki, S. (2000) The architecture of the colour centre in the human visual brain: new results and a review. *Eur. J. Neurosci.*, **12**, 172–193.

Bernardo, J.M. & Smith, A.F.M. (2008) *Bayesian Theory*, Bayesian Theory.

Bornstein, M.H., Kessen, W., & Weiskopf, S. (1976) The categories of hue in infancy. *Science*, **191**, 201–202.

Botvinick, M. & Toussaint, M. (2012) Planning as inference. *Trends Cogn. Sci.*,.

Bovens, L. & Hartmann, S. (2005) *Bayesian Epistemology*, Bayesian Epistemology.

Brouwer, G.J. & Heeger, D.J. (2013) Categorical clustering of the neural representation of color. *J. Neurosci.*, **33**, 15454–15465.

Bushnell, I.W.R., Sai, F., & Mullin, J.T. (1989) Neonatal recognition of the mother's face. *Br. J. Dev. Psychol.*, **7**, 3–15.

Chen, C.-H. & Zeki, S. (2011) Frontoparietal activation distinguishes face and space from artifact concepts. *J. Cogn. Neurosci.*, **23**, 2558–2568.

Clark, A. (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. BRAIN Sci. 36, 181–253*, 181–253.

Dayan, P., Hinton, G.E.E., Neal, R.M.M., & Zemel, R.S.S. (1995) The helmholtz machine. *Neural Comput.*, **7**, 889–904.

Fienberg, S.E. (2006) When did Bayesian inference become "Bayesian"? *Bayesian Anal.*, **1**, 1–40.

Friston, K., Mattout, J., & Kilner, J. (2011) Action understanding and active inference. *Biol. Cybern.*, **104**, 137–160.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., FitzGerald, T., & Pezzulo, G. (2015) Active inference and epistemic value. *Cogn. Neurosci.*, 150217111908007.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004) *Bayesian Data Analysis*, Chapman Texts in Statistical Science Series.

Gelman, A. & Shalizi, C.R. (2013) Philosophy and the practice of Bayesian statistics. *Br. J. Math. Stat. Psychol.*, **66**, 8–38.

Good, I.J., Hacking, I., Jeffrey, R.C., & Törnebohm, H. (1966) The Estimation of Probabilities: An Essay on Modern Bayesian Methods. *Synthese*, **16**, 234–244.

Goren, C.C., Sarty, M., & Wu, P.Y. (1975) Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics*, **56**, 544–549.

Grill-Spector, K. (2003) The neural basis of object perception. *Curr. Opin. Neurobiol.*, **13**, 159–166.

Haxby, J., Hoffman, E., & Gobbini, M. (2000) The distributed human neural system for face perception. *Trends Cogn. Sci.*, **4**, 223–233.

Helmholtz, H. von (2001) Concerning the perceptions in general. In *Visual Perception: Essential Readings*. pp. 24–44.

Hubel, D.H. & Wiesel, T.N. (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.*, **160**, 106–154.

Hubel, D.H. & Wiesel, T.N. (1977) Ferrier lecture. Functional architecture of macaque monkey visual cortex. *Proc. R. Soc. Lond. B. Biol. Sci.*, **198**, 1–59.

Ingle, D. (1985) The goldfish as a retinex animal. *Science (80-.).*, **227**, 651–654.

Ishai, A., Schmidt, C.F., & Boesiger, P. (2005) Face perception is mediated by a distributed cortical network. *Brain Res. Bull.*, **67**, 87–93.

Ishizu, T., Ayabe, T., & Kojima, S. (2008) Configurational factors in the perception

of faces and non-facial objects: an ERP study. *Int. J. Neurosci.*, **118**, 955–966.
Johnson, M.H., Dziurawiec, S., Ellis, H., & Morton, J. (1991) Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, **40**, 1–19.
Kant, I. (1781) *Kritik Der Reinen Vernunft, 1st Edition, Translated by WS Pluhar (1996) as Critique of Pure Reason*, 1st edn. Hackett, Indianapolis.
Kant, I. (1783) *Prologemena to Any Future Metaphysics (Translated by PG Lucas)*. Manchester University Press, Manchester.
Kant, I. (1787) *Kritik Der Reinen Vernunft, 2nd Edition, Translated as Critique of Pure Reason by WS Pluhar*. Hackett, Indianapolis.
Kanwisher, N., McDermott, J., & Chun, M.M. (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.*, **17**, 4302–4311.
Kersten, D., Mamassian, P., & Yuille, A. (2004) Object Perception as Bayesian Inference. *Annu. Rev. Psychol.*, **55**, 271–304.
Knill, D.C. & Pouget, A. (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS Neurosci.*, **27**, 712–719.
Lafer-Sousa, R, Conway, BR, Kanwisheer, N. (2016) Color-biased regions of the ventral visual pathway lie between face- and place-selective regions in humans, as in macaques. *J. Neurosci.*, **36**, 1682–1697.
Land, E. (1974) The retinex theory of colour vision. *Proc. R. Inst. G. B.*, **47**, 23–58.
Land, E. (1985) Statement mad by Land at a Stated Meeting of the American Academy of Arts & Sciences. *Stated Meet. Rep.*, 7–8.
Land, E.H. (1986) An alternative technique for the computation of the designator in the retinex theory of color vision. *Proc. Natl. Acad. Sci. U. S. A.*, **83**, 3078–3080.
Land, E.H. & McCann, J.J. (1971) Lightness and Retinex Theory. *J. Opt. Soc. Am.*, **61**, 1–11.
Lee, T.S. & Mumford, D. (2003) Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A. Opt. Image Sci. Vis.*, **20**, 1434–1448.
marquis de Laplace, P.S. (1812) *Théorie Analytique Des Probabilités*. V. Courcier.
Meadows, J.C. (1974) Disturbed perception of colours associated with localized cerebral lesions. *Brain*, **97**, 615–632.
Natu, V.S., Barnett, M.A., Hartley, J., Gomez, J., Stigliani, A., & Grill-Spector, K. (2016) Development of Neural Sensitivity to Face Identity Correlates with Perceptual Discriminability. *J. Neurosci.*, **36**, 10893–10907.
Peelen, M. V & Downing, P.E. (2007) The neural basis of visual body perception. *Nat. Rev. Neurosci.*, **8**, 636–648.
Pitcher, D. (2014) Facial Expression Recognition Takes Longer in the Posterior Superior Temporal Sulcus than in the Occipital Face Area. *J. Neurosci.*, **34**, 9173–9177.
Pouget, A., Beck, J.M., Ma, W.J., & Latham, P.E. (2013) Probabilistic brains: knowns and unknowns. *Nat. Neurosci.*, **16**, 1170–1178.
Pourtois, G., Schwartz, S., Seghier, M.L., Lazeyras, F., & Vuilleumier, P. (2005) View-independent coding of face identity in frontal and temporal cortices is modulated by familiarity: An event-related fMRI study. *Neuroimage*, **24**, 1214–1224.
Rao, R.P.N. & Ballard, D.H. (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.*, **2**, 79–87.

Rosenkrantz, R.D. (1977) *Inference, Method and Decision: Towards a Bayesian Philosophy of Science*. D. Reidel Publishing Company.

Schmid, M.C., Panagiotaropoulos, T., Augath, M.A., Logothetis, N.K., & Smirnakis, S.M. (2009) Visually driven activation in macaque areas V2 and V3 without input from the primary visual cortex. *PLoS One*, **4**, e5527.

Sergent, J., Ohta, S., & MacDonald, B. (1992) Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain*, **115**, 15–36.

Shigihara, Y. & Zeki, S. (2013) Parallelism in the brain's visual form system. *Eur. J. Neurosci.*, **38**, 3712–3720.

Shigihara, Y. & Zeki, S. (2014a) Parallel processing in the brain's visual form system: an fMRI study. *Front. Hum. Neurosci.*, **8**, 506 doi: 10.3389/fnhum.2014.00506.

Shigihara, Y. & Zeki, S. (2014b) Parallel processing of face and house stimuli by V1 and specialized visual areas: a magnetoencephalographic (MEG) study. *Front. Hum. Neurosci.*, **8**, 901:doi: 10.3389/fnhum.2014.00901.

Stoughton, C.M. & Conway, B.R. (2008) Neural basis for unique hues. *Curr. Biol.*, **18**, R698-9.

Talbott, W. (2008) Bayesian Epistemology. *Stanford Encycl. Philos.*,.

Yang, J., Kanazawa, S., Yamaguchi, M.K., & Kuriki, I. (2016) Cortical response to categorical color perception in infants investigated by near-infrared spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.*, **113**, 2370–2375.

Yuille, A. & Kersten, D. (2006) Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.*, **10**, 301–308.

Zeki, S. (1973) Colour coding in rhesus monkey prestriate cortex. *Brain Res.*, **53**, 422–427.

Zeki, S. (1978) Uniformity and diversity of structure and function in rhesus monkey prestriate visual cortex. *J. Physiol.*, **277**, 273–290.

Zeki, S. (1980) The representation of colours in the cerebral cortex. *Nature*, **284**, 412–418.

Zeki, S. (1983) The distribution of wavelength and orientation selective cells in different areas of monkey visual cortex. *Proc. R. Soc. London B*, **217**, 449–470.

Zeki, S. (1984) The construction of colours by the cerebral cortex. *Proc. R. Inst. G. B.*, **56**, 231–257.

Zeki, S. (1990) A century of cerebral achromatopsia. *Brain*, **113**, 1721–1777.

Zeki, S. (1993) *A Vision of the Brain*. Blackwell Scientific, Oxford.

Zeki, S. (1999) *Inner Vision: An Exploration of Art and the Brain.*, Optometry and Vision Science. Oxford University Press, Oxford.

Zeki, S. (2009) *Splendors and Miseries of the Brain: Love, Creativity and the Quest for Human Happiness*. Wiley-Blackwell, Oxford.

Zeki, S. (2013) Clive Bell's "Significant Form" and the neurobiology of aesthetics. *Front. Hum. Neurosci.*, **7**, 730.

Zeki, S. (2016) Multiple asynchronous dtimulus- and task-dependent hierarchies (STDH) within the brain's parallel processing systems. *Eur. J. Neurosci.*,.

Zeki, S. & Ishizu, T. (2013) The "Visual Shock" of Francis Bacon: an essay in neuroesthetics. *Front. Hum. Neurosci.*, **7**, 850.

Zeki, S., Watson, J., & Lueck, C. (1991) A direct demonstration of functional specialization in human visual cortex. *J. Neurosci.*,.

Zeki, S., Watson, J.D., Lueck, C.J., Friston, K.J., Kennard, C., & Frackowiak, R.S. (1991) A direct demonstration of functional specialization in human visual cortex. *J. Neurosci.*, **11**, 641–649.

Supplementary Materials

Part I:

Let $\mathfrak{C}$ be the posterior of *C* (colour category), $\mathfrak{H}$ the posterior of *H (hue)*, β^c and β^H the *β priors* for *C* and *H,* and ε depends on the experiment conducted. The following three steps constitute the theorem.

Step I (Biological β prior): inherited or rapidly acquired very shortly after birth, a *β prior* is neurobiologically immutable; it is constant throughout one's life span, and is almost totally resistant to culture and learning. Notationally, we can say that the posterior for colour, $\mathfrak{C}$, is deterministic of the *β prior* for the colour category, β^C, or

$$\mathfrak{C} \leftarrow \boldsymbol{\beta^C} \qquad (1)$$

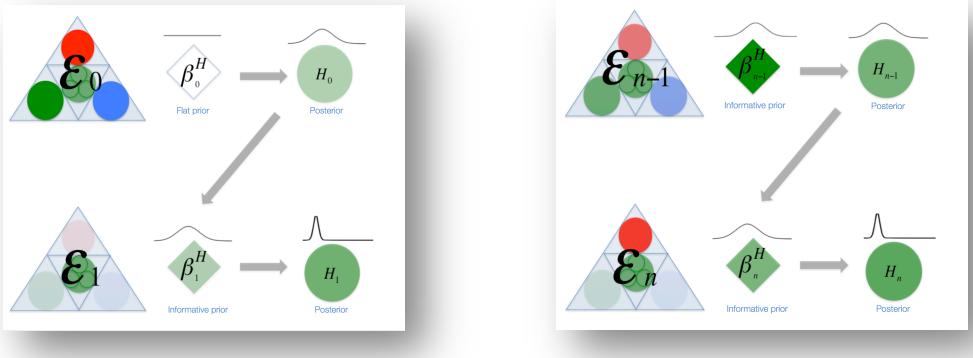
Step 2 (*Initialization*): Immediately after birth, the β prior is incorporated with the first scene ($\boldsymbol{\varepsilon_0}$) viewed along with a prior (derived from viewing a new scene containing a hue); next, the illuminants are adjusted so that the green patch is reflecting more red light and the shade of green changes to a darker green which becomes the initial posterior $\mathfrak{H}_0$. Notationally, we say the initial posterior for hue, $\boldsymbol{\mathfrak{H}_0}$, is dependent upon *β prior* β^H for hue and the first experiment conducted $\boldsymbol{\varepsilon_0}$, or

$$\boldsymbol{\mathfrak{H}_0} \propto \boldsymbol{\varepsilon_0}\, \boldsymbol{\beta_0^H} \qquad (2)$$

Step 3 (*Adaptive Learning*): The illuminants are re-adjusted (the same scene is viewed in a different illuminant) so that the green patch is now reflecting more green light. Now its hue changes to a much brighter green. The posterior formed in Step 2 then becomes a new (now more informative) *β prior* of hue (call it β_1^H) and along with a new scene ($\boldsymbol{\varepsilon_1}$), forms a new posterior of hue $\boldsymbol{\mathfrak{H}_1}$. This process continues throughout the entire life span: whenever new experiments occur, old posteriors become the new *priors*, and together they form the new posteriors, Notationally, for the n^{th} iteration:

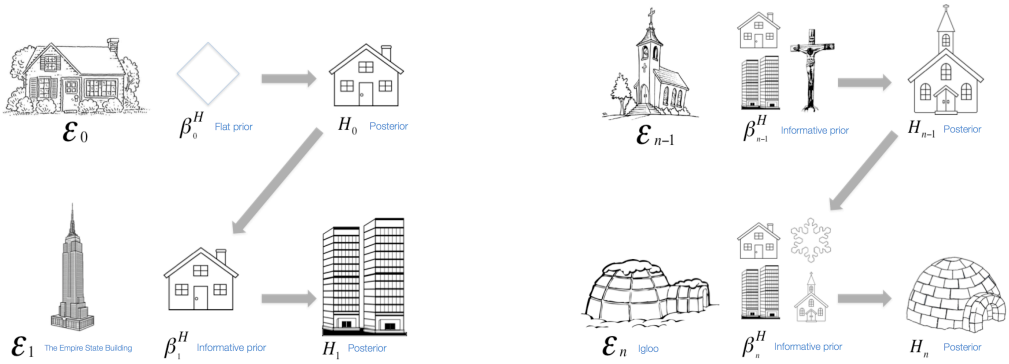
$$\boldsymbol{\mathfrak{H}_n} \propto \boldsymbol{\varepsilon_n}\, \boldsymbol{\beta_n^H} \qquad \mathbf{(2)}$$

The procedure is illustrated in Figure 3 below



(a) Initialization (b) Iterative update

Figure 3: The Bayesian brain procedures (a) once an experiment is conducted (for example by viewing a patch of colour in different illuminants), the brain incorporates the experimental data ε and the prior regarding a quantity of interest (e.g. hue) and thus forms the posterior for that quantity. In the absence of any prior regarding that quantity (e.g. when the observer has never seen a dark green hue before the experiment), the brain naturally assigns a non-informative *prior* (e.g. a flat prior β_0^H); if the observer had experienced that hue before the experiment, the brain assigns an informative *prior* (e.g. a Gaussian density with high probability around dark green, or β_1^H); (b) once the posterior (e.g. H_0) is formed, it then (possibly instantaneously) becomes the new *prior* (i.e. β_1^H), that, together with a new experiment (e.g. ε_1), forms a new posterior (H_1). This process continues throughout one's life.



(a) Initialization (b) Iterative update

Figure 4: An Example of the Bayesian Brain Procedures in Viewing Houses. (a) at the top, a simple house is viewed (e.g. ε_0). In the absence of any prior regarding houses, a *noinformative prior* (e.g. a flat prior β_0^H) is incorporated, resulting in an initial posterior. Note that this β_0^H, despite being noninformative, is dependent upon the brain's biological prior, for example, for angles and lines. Once a posterior (e.g. H_0) is formed, it then (possibly instantaneously) becomes the new *prior* for houses (i.e. β_1^H), that, together with viewing a new house (now the Empire State Building, e.g. ε_1), forms a new posterior for houses (H_1). (b) This process continues throughout one's life. When a church (e.g. ε_{n-1}) is viewed, one incorporates priors for houses and for a cross (e.g. β_{n-1}^H above), and forms a new posterior for houses(H_{n-1}). Similarly, when an lgloo (e.g. ε_n) is viewed, one incorporates priors for houses and for snow (e.g. β_n^H above), and forms a new posterior for houses (H_n). In general, the order of which houses are viewed does not matter as much in forming posteriors. Yet, throughout our life, we tend to encounter close to our culture first, then extend to complex houses and houses of exotic cultures. The above figure summarizes such an experience.

Supplementary Materials

Part II

Proof of Equations (1) and (2):

For simplicity, let us prove Equations (1) and (2) in the context of colour vision.

First, let us define *H, C, E* as hue, colour category, and the experiment conducted (where we change O and S, wavelength composition from the centre, and wavelength composition from the surrounds, respectively). Let us further define *R* as the ratio in the amount of light of different wavebands reflected from the centre and from the surrounds (hereinafter ratio).

Given *O, S*, the ratio *R* that forms colour category is fixed:

$$\boldsymbol{R \equiv O/S}\ ^{5}.$$

By *Bayes'* rule, we have

$$[\mathbf{C|E, R}] \propto [\boldsymbol{E|C, R}][\boldsymbol{C|R}][\boldsymbol{R}] \propto [\boldsymbol{R}]$$

where the second $\propto$ follows from: since C is deterministic of R, then $\mathbb{P}$(C|R) is either 1 or 0, and that C and R are independent of E. Denoting [C|E, R] as the posterior $\mathfrak{C}$ for colour category and $[R]$ as the β prior β^C for colour category, we have proven equation (1).

Next,

$$[\boldsymbol{H|C, E, R}] \propto [\boldsymbol{E|H, C, R}]\,[\boldsymbol{H|C, R}][\boldsymbol{R|C}][\boldsymbol{C}] \propto [\boldsymbol{E|H}][\boldsymbol{H|C, R}][\boldsymbol{R}]$$

where the second $\propto$ follows from: C and R are independent of E, P(C|R) is either 1 or 0, and C is deterministic of R. Denoting [H|C, E, R] as the posterior $\mathfrak{H}$ for hue, $[\boldsymbol{H|C, R}][\boldsymbol{R}]$ as the β prior β^H for hue, and $[E|H]$ as ε the experiment conducted, we have $\mathfrak{H} \propto \varepsilon\, \beta^H$, which proves equation (2).

Remarks:

Note that the β *prior* β^H for hue constitutes two parts $[\boldsymbol{H|C, R}]$ an informative prior for hue, and $[\boldsymbol{R}]$, the β prior β^C for colour category.

---

[5] Here, O and S are both three-dimensional. For example, 0 = (g^{l_1}, g^{m_1}, and g^{s_1}) and S = (g^{l_2}, g^{m_2}, and g^{s_2}). See Section III B of the article for further details.