

The Bayesian brain: What is it and do humans have it?

Dobromir Rahnev 

School of Psychology, Georgia Institute of Technology, Atlanta, GA 30332.

rahnev@psych.gatech.edu

www.rahnevlab.gatech.edu

doi:10.1017/S0140525X19001377, e238

Abstract

It has been widely asserted that humans have a “Bayesian brain.” Surprisingly, however, this term has never been defined and appears to be used differently by different authors. I argue that Bayesian brain should be used to denote the realist view that brains are actual Bayesian machines and point out that there is currently no evidence for such a claim.

In his target article, Brette criticized the claim that people have a “Bayesian brain.” This term has been widely adopted to describe the nature of the human brain (Friston 2012; Knill & Pouget 2004; Sanborn and Chater 2016). Surprisingly, however, there is no agreed-upon definition of the term. Two rather informal definitions have been offered. First, Knill and Pouget (2004) describe the “Bayesian coding hypothesis” as follows: “the brain represents sensory information probabilistically, in the form of probability distributions”; second, according to Friston (2012), the “Bayesian brain says that we are trying to infer the causes of our sensations based on a generative model of the world.” Neither of these definitions even mentions Bayesian computations, which, one may expect, should be central to the idea of a Bayesian brain. So, what then, is exactly meant by the “Bayesian brain?”

Any model of Bayesian computation contains at a minimum a set S of known stimuli, a set r of internal responses, and a known generative model $P(r|S)$ of the response generated by each stimulus. Bayes’ theorem is used to invert the generative model to compute a likelihood function that is then combined with a prior $P(S)$ to obtain a posterior distribution. The result can be used to inform a forthcoming action or simply the percept of the observer.

A *Bayesian* brain must be implementing such Bayesian computations on some level. One can distinguish between two possible views here (Block 2018). The “as if” view holds that the brain does not necessarily literally have a generative model and does not literally use Bayes’ theorem to derive a likelihood function. Instead, the computations performed by the brain can be seen “as if” it performs these operations. The “realist” view, on the other hand, holds that a generative model, a likelihood function, and a prior are actually represented in the brain and that the computations performed are literally the computations required by Bayes’ theorem. Unfortunately, most authors do not necessarily commit to one or the other interpretation and, in some cases, appear to make different theoretical commitments in different papers.

Importantly, the “as if” view is typically expressed at Marr’s “computational level” with no commitment to brain implementation (Griffiths et al. 2012). Consequently, using the term “Bayesian

brain” in an “as if” sense appears almost contradictory because this usage is explicitly *not* about what happens in the brain. Thus, if the “Bayesian *brain*” is really a claim about the brain, then it has to be reserved for the realist view that the brain literally implements the components of Bayesian computation.

Is there evidence for the claim that humans have a Bayesian brain in the realist sense? No direct evidence has been presented to date. Instead, what is usually offered is an indirect argument from behavior. For example, Knill and Pouget (2004) motivated the view that brains are Bayesian by “the myriad ways in which human observers behave as optimal Bayesian observers” (p. 712). The problem is that this argument ignores the fact that findings of suboptimality are at least as common as findings of optimality (Rahnev and Denison 2018). Even more importantly, Bayesian optimality can be achieved by non-Bayesian algorithms (Ma 2012), and thus, such findings do not imply that brain computations are literally Bayesian.

In fact, as Brette eloquently explains, there are many reasons to doubt that brains are literally implementing Bayesian computations. Here, I formalize some of the issues examined by Brette and discuss some additional problems.

First, as pointed out by Brette, the internal response depends on more than just the stimulus of interest. Instead, the internal response to, for example, a tilted bar is better described not as $P(r|S)$ but as $P(r|S, \Theta)$, where Θ is a set of variables that affect neural firing, including the color of the bar, the color of the background, the size of the bar, the level of illumination, contrast, attention, arousal, metabolic state, and so forth. Dozens of such “confounding” variables can easily be present in any real-world situation. Inverting this generative model necessitates the integration (i.e., marginalization) over all possible values of all of these variables. For many forms of the assumed internal response, this computation is infeasible in real brains.

Second, as also discussed by Brette, Bayesian computations depend on the existence of a well-defined response r . However, brain activity is a dynamic, recurrent, never-ending string of action potentials. It is unclear how the Bayesian brain isolates “the response” to any given stimulus to perform the necessary Bayesian computations.


Third, an even more insidious problem that Brette did not examine in the context of the Bayesian brain is that a realist Bayesian brain must already know the set S of possible stimuli and the generative model $P(r|S)$ for each stimulus. However, the brain has to first learn both the stimuli in the world and their associated generative models. A truly Bayesian brain would thus form a probability distribution over the stimuli and generative models, which goes against current models that assume the existence of a predefined set S of stimuli.

Finally, a central tenet of the Bayesian brain – that the brain represents and computes with full probability distributions – has only been supported by theoretical proposals of how this *could* be achieved. Recent empirical research has, however, challenged this tenet (Yeon and Rahnev 2019).

The idea of the “Bayesian brain” has gained popularity perhaps not despite but because of the fact that it has never been clearly defined. This ambiguity shields it from criticism but it also robs it from any chance of contributing to scientific progress. To be useful, the term should be defined according to its plain meaning of a realist view where the brain literally represents the different components of Bayesian computations and researchers should present evidence for it that goes beyond “some behavior is close

to optimal.” Until then, the “Bayesian brain” should be seen for what it is: a theoretical possibility fully divorced and shielded from the empirical reality.

Not just a bad metaphor, but a little piece of a big bad metaphor

George N. Reeke Jr. 

Laboratory of Biological Modeling, The Rockefeller University, New York, NY 10065.

reeke@rockefeller.edu

<https://www.rockefeller.edu/our-scientists/heads-of-laboratories/892-george-n-reeke-jr/>

doi:10.1017/S0140525X19001225, e239

Abstract

Besides failing for the reasons Brette gives, codes fail to help us understand brain function because codes imply algorithms that compute outputs without reference to the signals' meanings. Algorithms cannot be found in the brain, only manipulations that operate on meaningful signals and that cannot be described as computations, that is, sequences of predefined operations.

Brette finds fault with the coding metaphor for neuronal activity in the brain on the basis of its disconnection with the causal structure of brain activity and its inadequate representational power. In so doing, he shows why brain activity is not compatible with a computational picture that includes coding of sensory signals, computation with those codes, and then decoding to generate behavior. The quotations in Section 4.1 even suggest that decoding must occur before the brain can interpret codes to determine action. Only in one sentence in his final section, 5.2 (para. 2), does he come near to noticing the real problem with coding: “Even if it were possible to map brain activity to computational descriptions, neural codes would not provide the adequate mapping.” He is correct about adequate mappings, but the bigger problem is the one implicit in the “even if” clause: Computational descriptions are not the way to describe what it is the brain does.

First let me clear away one objection to my argument: Yes, the brain computes if we look upon it as a device that receives sensory signals encoded as neuronal firings and emits behavioral commands also encoded as neuronal firings. I think it is useful to constrain “computation” to its nonmetaphorical usage to describe what goes on in Turing or Von Neumann computers – not to be a stickler for definition, but because the aspects in which the activity in the brain differs from the activity in those machines are precisely the things that are at the heart of the hard problems of neuroscience, the things that the computational metaphor drives researchers to look for that are not there: meanings assignable externally to neuronal firings and algorithms that describe a finite sequence of steps to get from a defined input to a defined output, that is, programs. Without externally assigned meanings and programs to operate upon them, computation is only a metaphor, in my view a big bad metaphor that has only held back the science of the brain.

Why is the computer metaphor bad? Because it inspires people to look for codes and algorithms as solutions to these basic problems instead of looking for mechanisms relevant to the brain. For example, it led Tsotsos (2011) to the absurd, admittedly strawman, conclusion that a general unbounded visual match on an image with p pixels requires time on order $O(p^2 2^p)$. So the brain must be doing something else. Perhaps rather than search for visual algorithms, one could address questions like these: How does the firing of a neuron in the brain come to signify something to those neurons that receive that firing, as opposed to signifying something to the experimenter who records it? How do these firings organize themselves, as a result of experience in the world, to produce behavioral outputs that serve the survival needs of the organism, *without* an external programmer?

As Brette is well aware, the meaning of a neuronal spike, unlike a bit in a computer, cannot be described in isolation. Perhaps the best discussions of how neural firings come to have significance for other neurons are those provided by Harnad (1990a) and Bickhard & Terveen (1996). It won't do just to add more codes (Brette, sect. 1, last para.). Neurons are members of assemblies that form and re-form according to the situation (Izhikevich 2006); the meaning of neuronal firing depends on context (Gilbert 1996) and may differ for different recipient neurons. Analyzing neuronal firing from the point of view of an “ideal observer” is useless because neuronal firing is not just a well-defined but noisy code; for success one needs a more complicated observer, perhaps a “homunculus,” which can vary in its responses according to the total picture provided by all the other neurons in the system, interacting through their recurrent or reentrant connections. But then one has just kicked the can down the road; such a homunculus is not a computer. It is not fair to silently ascribe key elements of the performance of a brain model to components that are not included in the model, the unmodeled homunculus in the machine (Reeke and Edelman 1988).

In short, coding fails because the only thing it is good for is as input to and output from algorithms. But if not algorithms, then what? The standard computer science picture of algorithms, even including those that emulate nondeterministic physical phenomena, is still the Turing machine definition: a predefined sequence of operations taken from a predefined set designed to accomplish a predefined computation. With this broad definition, algorithms can no doubt be found in the brain. But what are the predefined operations: membrane depolarization, spike firing, volume diffusion of chemical signals? How are these operations organized without a programmer: synaptic plasticity regulated by multiple chemical signals conveying states of arousal, emotions, homeostasis, reward, and punishment? And what are the predefined computations, or effective methods of performing behaviors: obtaining food, water, mates, or just some ill-defined pleasure signal? The answers to these questions are not found in algorithm theory.

Fodor and Pylyshyn (1988) have argued persuasively that so-called connectionist models (Rumelhart et al. 1986) are not sufficient to implement all cognitive activities of brains; symbol systems and syntactic operations on them are needed. There is no contradiction once one looks at real brains: symbol systems and syntactic operations upon them can be constructed from the signals and operations upon them that I have just argued we need to look for in the brain. The question we only have partial answers to is how this is accomplished by experience in the complex real world. Computation theory does not provide the answer to that problem. Brette's final suggestion, that the solution resides somewhere in the area of modeling the full sensorimotor loop, is