Behavioral and Brain Sciences

cambridge.org/bbs

Is coding a relevant metaphor for the brain?

Romain Brette

Institut de la Vision, Université Pierre-and-Marie-Curie 06, Sorbonne Universités, INSERM, CNRS, 75012 Paris, France
romain.brette@inserm.fr http://romainbrette.fr

Target Article

Cite this article: Brette R. (2019) Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences* **42**, e215: 1–58. doi:10.1017/S0140525X19000049

Target Article Accepted: 5 September 2017
Target Article Manuscript Online: 16 July 2018
Commentaries Accepted: 14 January 2019

Keywords:
action; information; neural coding; perception; sensorimotor

What is Open Peer Commentary? What follows on these pages is known as a Treatment, in which a significant and controversial Target Article is published along with Commentaries (p. 14) and an Author's Response (p. 44). See bbsonline.org for more information.

Abstract

"Neural coding" is a popular metaphor in neuroscience, where objective properties of the world are communicated to the brain in the form of spikes. Here I argue that this metaphor is often inappropriate and misleading. First, when neurons are said to encode experimental parameters, the neural code depends on experimental details that are not carried by the coding variable (e.g., the spike count). Thus, the representational power of neural codes is much more limited than generally implied. Second, neural codes carry information only by reference to things with known meaning. In contrast, perceptual systems must build information from relations between sensory signals and actions, forming an internal model. Neural codes are inadequate for this purpose because they are unstructured and therefore unable to represent relations. Third, coding variables are observables tied to the temporality of experiments, whereas spikes are timed actions that mediate coupling in a distributed dynamical system. The coding metaphor tries to fit the dynamic, circular, and distributed causal structure of the brain into a linear chain of transformations between observables, but the two causal structures are incongruent. I conclude that the neural coding metaphor cannot provide a valid basis for theories of brain function, because it is incompatible with both the causal structure of the brain and the representational requirements of cognition.

1. Introduction

A pervasive paradigm in neuroscience is the concept of neural coding (deCharms and Zador 2000): the query "neural coding" on Google Scholar retrieved about 15,000 papers in the last 10 years. Neural coding is a communication metaphor. An example is the Morse code (Fig. 1A), which was used to transmit texts over telegraph lines: each letter is mapped to a binary sequence (dots and dashes). In analogy, visual signals are encoded into the spike trains of retinal ganglion cells (Fig. 1B). Both the Morse code and the retinal code relate to a communication problem: to communicate text messages over telegraph lines, or to communicate visual signals from the eye to the brain. This problem has been formalized by communication theory (Shannon 1948), also called information theory, a popular tool in neuroscience (Rieke et al. 1997).

The neural coding metaphor has shaped neuroscience thinking for more than five decades. Barlow (1961) used the metaphor extensively in his work on sensory neurons, although he warned to "not regard these ideas as moulds into which all experimental facts must be forced." In a seminal review entitled *Neural Coding*, Perkel and Bullock (1968) depicted "the nervous system [as] a communication machine" and already recognized the "widespread use of "code" in neuroscience." An illustration of hieroglyphs figures prominently at the top of the technical appendix. Around the same time, entire books were devoted to "sensory coding" (Somjen 1972; Uttal 1973).

As the linguists Lakoff and Johnson (1980a) have argued, the metaphors that pervade our language are not neutral; on the contrary, they form the architecture of our conceptual system. What are the concepts carried by the neural coding metaphor that make it a possibly relevant metaphor for the activity of the retina? There are three key properties (Fig. 1C), which are all used in Perkel and Bullock's (1968) review:

1. The technical sense of a code is a correspondence between two domains, for example, visual signals and spike trains. We call this relation a *code* to mean that spike trains specify the visual signals, as in a cipher: one can theoretically reconstruct the original message (visual signals) from the encoded message (spike trains) with some accuracy, a process called *decoding*. Information theory focuses on statistical aspects of this correspondence (Shannon 1948). It is in this sense that neurons in the primary visual cortex encode the orientation of bars in their firing rate, neurons in the auditory brainstem encode the spatial position of sounds (Ashida and Carr 2011), and neurons in the hippocampus encode the animal's location (Moser et al. 2008).

© Cambridge University Press 2019

CAMBRIDGE UNIVERSITY PRESS

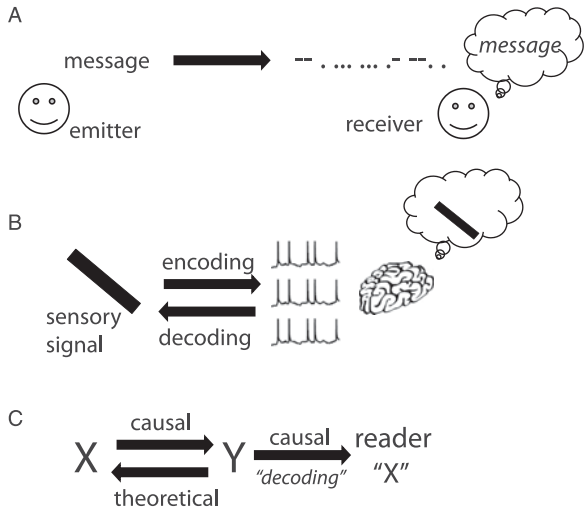


Figure 1. The coding metaphor. (A) An emitter transmits a message to a receiver in an altered form called "code" (here the Morse code). The receiver knows the correspondence and can reconstruct ("decode") the original message. (B) In analogy, visual signals are encoded in the spike trains of the optic nerve. The rest of the visual system treats these spike trains as visual information. (C) Implicit structure of the neural coding metaphor ("*Y* encodes *X*"). There is correspondence between *X* and *Y*. Encoding refers to a causal mechanism from *X* to *Y*, and decoding is a theoretical inverse mapping; *Y* causes changes in the reader (often improperly called "decoding") and represents *X* in some sense.

2. Yet, not all cases of correlations in nature are considered instances of coding. Climate scientists, for example, rarely ask how rain encodes atmospheric pressure. Another key element of the coding metaphor is that the spike trains are considered messages for a reader, the brain, about the original message: this is the representational sense of the metaphor. Perkel and Bullock call the reader's activity "interpretation of the encoded information." In his book on sensory coding, Somjen (1972) writes: "Information that has been coded must at some point be decoded also; one suspects, then, that somewhere within the nervous system there is another interface … where 'code' becomes 'image.'" Similar statements abound in modern neuroscience literature: "A stimulus activates a population of neurons in various areas of the brain. To guide behavior, the brain must correctly decode this population response and extract the sensory information as reliably as possible" (Jazayeri and Movshon 2006).
3. Finally, we would not say that visual signals encode retinal spike trains, even though this would comply with the technical sense. The reason is that the communication metaphor implicitly assumes a causal relation between the original message and the encoded message; here, spike trains result from visual signals by a causal process (transduction). Similarly, to be a representation for a reader, the neural code must at least have a causal effect on the reader. This causal structure is implicit in Perkel and Bullock's (1968) definition of neural coding: "the transformations of information in the nervous system, from receptors through internuncials to motor neurons to effectors."

Romain Brette is a theoretical neuroscientist at the Vision Institute, Paris. He was previously Assistant Professor at the Departments of Computer Science and Cognitive Science of Ecole Normale Supérieure, Paris. He has authored over 65 articles on various topics in neuroscience, including auditory perception, for which he was awarded the early career scientific prize from Fondation pour l'Audition. His current work focuses on the development of integrative models of perception, emphasizing the interaction of the organism with the environment.

These three elements (correspondence, representation, causality) constitute the conceptual scaffold of the neural coding metaphor. It could be argued that most technical work on neural coding uses only the first technical sense (correspondence), where the word *code* is used as a synonym for *correlate*. The use of the metaphor would thus amount to only an inappropriate but innocuous choice of words. But what is the scope of neural codes if they have no causal powers? In his famous critique of Skinner's behaviorism, Chomsky et al. (1959) summarizes the problem with the improper use of metaphors: "[Skinner] utilizes the experimental results as evidence for the scientific character of his system of behavior, and analogic guesses (formulated in terms of a metaphoric extension of the technical vocabulary of the laboratory) as evidence for its scope." The goal of this article is to demonstrate that this quote fully applies to the neural coding metaphor, where "scope" is a particular theory of brain function implied by the conceptual structure of the metaphor.

The general argument is as follows. Scientific claims based on neural coding rely on the representational sense or at least on the causal sense of the metaphor. But none of these two senses is implied by the technical sense (correspondence). When we examine the representational power of neural codes (part 1), we realize that coding variables are shown to correlate with stimulus properties but the code depends on the experimental context (stimulus properties, protocol, etc.). Therefore, neural codes do not provide context-free symbols. But context cannot be provided by extending the code to represent a larger set of properties, because context is what defines properties (e.g., the orientation of a bar). Thus, neural codes have little representational power. The fundamental reason (part 2) is that the coding metaphor conveys an inappropriate concept of information and representation (Bickhard 2009; Bickhard and Terveen 1996). Neural codes carry information by reference to things with known meaning. In contrast, perceptual systems have no other option than to build information from relations between sensory signals and actions, forming a structured internal model. Finally (part 3), the neural coding metaphor tries to fit the causal structure of the brain (dynamic, circular, distributed) into the causal structure of neural codes (atemporal, linear), substituting the arbitrary temporality of algorithms for the temporality of the underlying physical system. The two causal structures are incongruent. Without denying the usefulness of information theory as a technical tool, I conclude that the neural coding metaphor cannot constitute a valid basis for theories of brain function because it is disconnected from the causal structure of the brain and incompatible with the representational requirements of cognition.

2. Encoding stimulus properties

2.1. Encoding an experimental parameter

The activity of neurons is often said to encode properties, for example, "Many cortical neurons encode variables in the external world via bell-shaped tuning curves" (Seriès et al. 2004). Here the

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

authors refer to a particular type of experiment, where a parameterized stimulus is presented to an animal and the activity of a neuron is recorded. For example, the orientation of a small bar is varied and the activity of a neuron in the primary visual cortex is recorded (Fig. 1B). It is found that orientation and neural activity co-vary and, hence, the neuron's firing rate encodes the orientation of the bar in the sense of correspondence. What is the scope of such a proposition?

I will discuss a cartoon example from color perception, used by Francis Crick to warn against the "fallacy of the overwise neuron" (Crick 1979). Cones are broadly tuned to wavelength (Schnapf et al. 1987); in an experiment where light of different wavelengths is flashed, the amplitude of the transduced current varies systematically with wavelength (Fig. 2A). Thus, the current encodes wavelength in the technical sense of correspondence; one can recover wavelength from the magnitude of the current. Yet animals or humans with a single functional type of cone are color blind. Why are they color blind if their cones encode color information? This is clear in Figure 2A; if the current also depends on light intensity, then it does not provide unambiguous information about wavelength. In other words, the cone does not in fact encode wavelength in any general setting, even in the narrow sense of correspondence. The same remark applies to any tuning curve experiment.

Formally, the logical problem can be analyzed as follows. The tuning curve experiment shows a correspondence between stimulus parameter and current. This correspondence is composed of two parts (Fig. 2B): a mapping from wavelength to stimulus, which is experiment specific, and the transduction of stimulus into current. Thus, the experimental design ensures that there exists a mapping from wavelength to current. In other words, the proposition that the neuron encodes the experimental parameter is mainly a property of the experimental design rather than a property of the neuron (which only needs to be sensitive to the parameter). However, the situation is completely different in the real world, which is not constrained by the experimental design (Fig. 2C). In general, there might be a variety of stimuli, one of their properties being wavelength. In this case, there is a mapping from stimulus to wavelength and a mapping from stimulus to current, and it is not obvious at all that there is a mapping from wavelength to current, because current depends also on other properties. In this context, the proposition that the neuron encodes wavelength is a much stronger claim, but it is not at all entailed by the tuning curve experiment.[1] This confusion underlies influential neural coding theories of perception, for example, Bayesian theories (Jazayeri and Movshon 2006; Pouget et al. 2003), in which a neuron's firing rate is assumed to be a function of the stimulus parameter, rather than a context-dependent correlate (see sect. 2.3.2).

Thus, the correct interpretation of the tuning curve experiment is that the neuron is sensitive to the stimulus parameter, while to encode a property of stimuli (a "variable in the external world") is a somewhat orthogonal proposition; it means that the observable is not sensitive to other properties. For example, a color scientist would point out that wavelength is indeed not encoded by single cones, but by the relative activity of cones with different tunings (Fig. 2D), because that quantity does not depend on light intensity. Thus, referring to tuning curve experiments in terms of coding promotes a semantic drift, from the modest claim that a neuron is sensitive to some experimental manipulation to a much stronger claim about the intrinsic representational content of the neuron's activity. We will now see that this semantic drift indeed operates in current theories of brain function.

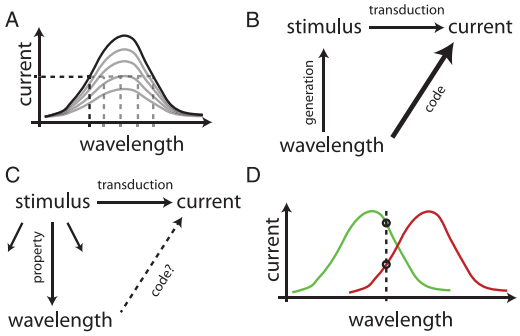


Figure 2. Encoding wavelength of light. (A) Response of a cone to flashed light as a function of wavelength (cartoon), at different intensities (grays). If intensity is fixed, wavelength can be inferred from transduced current. Otherwise, current is not informative about wavelength. (B) In a tuning curve experiment, the coding relation is implied by the experimental design: wavelength is mapped to stimulus, which is transduced into current. (C) If wavelength is only one property of a larger set of stimuli, there might be no coding relation. (D) The relative response of cones with different tunings may provide intensity-invariant information about wavelength.

2.2. The overwise neuron and its ideal observer

To understand how the neural coding metaphor unfolds, I will discuss one particular example in detail (but another one could have been chosen). In mammals, the major cue for sound localization in the horizontal plane is the difference in arrival times of the sound wave at the two ears (interaural time difference [ITD]) (Fig. 3A). Neurons in the medial superior olive (MSO) (in the

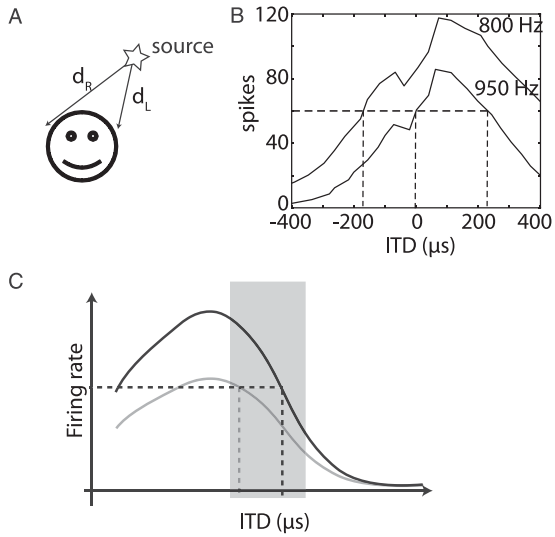


Figure 3. Encoding sound location. (A) A major cue for sound localization is the interaural time difference (ITD), $d_R - d_L$. (B) Number of spikes in response to two binaural tones (950 and 800 Hz) as a function of ITD for the same neuron in the medial superior olive of a cat (digitized from Yin and Chan [1990], Fig. 10). It is possible to infer the ITD from spike count if the experimental configuration (presented tone) is known, not if the sound is a priori unknown. (C) If the organism lived in a world with a single sound played at different ITDs, then the best way to encode ITD would be with a neuron tuned to an ITD outside the range of natural sounds (*shaded*), so that the selectivity curve is steep inside that range. However, the response of a single neuron is fundamentally ambiguous when sounds are diverse, irrespective of the steepness of the curve (selectivity curve for another sound shown in *gray*).

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

auditory brainstem) are sensitive to this cue (Joris et al. 1998); when a sound is played through earphones and the ITD is varied, the firing rate of those neurons changes (Fig. 3B). These neurons project to neurons in the inferior colliculus (IC), which also have diverse ITD tuning properties. Electrical stimulation in the cat's IC triggers an orienting response toward a particular contralateral direction, with stronger stimulations resulting in responses with a larger part of the body (one pinna, both pinnae, and eyes), by a pathway involving the superior colliculus (Syka and Straschill 1970). Unilateral lesions in the MSO or IC result in sound localization deficits in the contralateral field (Jenkins and Masterton 1982). Therefore, neurons in the IC have a critical role in localizing sounds in the contralateral field.

How does the activity of these neurons contribute to sound localization behavior? One way is consider the entire pathway, try to build a model of how neuron responses in various structures combine to produce an orientation reflex to a localized sound, and compare with the diverse experimental observations mentioned above. Another way is to ask how neurons encode sound location (McAlpine et al. 2001). It has, for example, been claimed that "there is sufficient information in the firing rates of individual neurons to produce ITD just-noticeable-differences that are comparable with those of humans psychophysically" (Shackleton et al. 2003, p. 723; Skottun 1998). What does this mean, and how significant is this fact?

The neuron of Figure 3B encodes ITD in the technical sense that one can estimate the ITD with some accuracy from the observation of the number of spikes, by inverting the tuning curve (i.e., decoding the neuron's response). It turns out that this accuracy is similar to the accuracy of sound localization by the animal. But the neuron's response is also sensitive to various aspects of sound (e.g., frequency, intensity), so our decoder would give totally inaccurate results in any other context. Therefore, the performance of this decoder is unrelated to our general ability to localize sounds. Yet, although the problem of ambiguities was acknowledged, it was concluded that "it might not be necessary to pool the outputs from many neurons to account for the high accuracy with which human observers can localize sounds" (Skottun 1998). This conclusion is unwarranted because tuning curves address the exactly orthogonal problem (sensitivity to ITD vs. insensitivity to other dimensions).

This incorrect conclusion is about localization, but what about discrimination? The first quote compared the tuning curve to discrimination performance, that is, psychophysical measurements of the ability to discriminate between two sounds that differ only by their ITD. This is a more restricted situation, but how can the responses of a single neuron be compared with the behavior of an organism without making any reference to the mechanisms that might link this neuron's activity to behavior (e.g., the pathway mentioned earlier)? More generally, how can a neural code be about behavior, when it is technically only about stimulus-response properties? This requires what Teller (1984) called a "linking proposition," an implicit postulate that directly relates neural activity to behavior. The linking proposition here, as in many neural coding studies, is that the brain implements an "ideal observer" (Macmillan and Creelman 2005). This is the representational sense of the metaphor, namely, the idea that neural responses are messages for a reader. The empirical question, then, is how plausible is this linking proposition?

Let us spell it out. The ideal observer reads the activity of the neuron. When the first stimulus is presented, it stores the number of spikes produced by the neuron in a window of a given duration (chosen by the experimenter) after the stimulus. It ignores all spikes produced before and after that window until the second stimulus is presented, upon which it stores again the number of spikes produced by the neuron. Then it retrieves the two stored numbers, compares them (and not others, e.g., the activity of other neurons), and decides to push one of two buttons. It is not so obvious how this ideal observer can be mapped to the pathway described above, for example, how the number of spikes of an arbitrary neuron in the brainstem, produced during several predefined time windows, can be stored in working memory for later comparison.

The ideal observer is ideal in the sense that it makes the best use of all available information. This includes the neural activity itself, but most importantly all the information that is available to the experimenter: when exactly the activity corresponds to the stimulus, what stimulus has been presented, the knowledge that the exact same sounds are played twice, which parts of the activity should be stored. On the other hand, the observer is not ideal in the sense that it uses nothing *more* than the information available to the experimenter. For example, if it also used the information available in other neurons (not recorded), then discrimination performance would be much better than psychophysical measurements. In other words, the ideal observer is not the best thing that the brain can do; it is the best thing that the experimenter can do.

Thus, by implying that the brain reads the neural code, we manage to make claims about perception and behavior while totally ignoring the mechanisms by which behavior is produced, as well as the constraints that the organism must face in ecological situations (e.g., not knowing the sound presentation protocol in advance). These claims rely on implicit linking propositions based on abstract constructs, where neural activity is likened to a processor register that the brain manages to store, retrieve, and manipulate, wherever it is in the brain and whenever it occurs. It would seem that empirical evidence or argumentation should be required to support such questionable hypotheses, since all conclusions are based on them. Why is it that no such justification is ever provided when "ideal observers" are introduced? The reason, it seems, is the semantic drift from the technical sense of code to the representational sense of code, which is logically flawed. The same flaw appears to underlie leading theories of neural population coding.

2.3. Populations of overwise neurons

2.3.1. Slope coding

What is the optimal way to encode ITD in the activity of neural populations? If all confounding dimensions (level, frequency, etc.) are neglected, then the best way to encode ITD is to have a steep monotonous relation between ITD and firing rate, that is, to maximize neural sensitivity to the ITD (Fig. 3C). Thus, the neuron's preferred ITD should lie outside the range of natural sounds (around ±800 μs for humans [Benichoux et al. 2016]) while the steepest slope of the selectivity curve should be inside. This is the concept of "slope coding." Thus, it has been argued that the optimal way to encode ITD is with two homogeneous populations of neurons with symmetrical tuning curves, peaking at ITDs that are not normally experienced (Harper and McAlpine 2004). Unfortunately, this conclusion is based entirely on the fallacy of the overwise neuron. If confounding dimensions are not neglected, then the opposite conclusion follows: As in the case of cones, heterogeneity of ITD tunings is crucial to resolve

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

the ambiguities resulting from nonspatial dimensions of sounds (Brette 2010; Goodman et al. 2013).

On the basis of the slope coding idea, a leading theory of sound localization (Grothe et al. 2010) proposes that sound location is encoded in the relative average activity of the two populations of neurons. It was initially meant to explain why many neurons are tuned to large ITDs that are not normally experienced[2] (McAlpine et al. 2001). Although using the relative activity of two populations somewhat reduces the ambiguity resulting from sound level, the fact that sounds have more than two dimensions again means that this model is unlikely to work in practice unless the auditory world consists of pure tones (Goodman et al. 2013). In any case, what needs to be demonstrated to support this theory is not that tuning curves have a steep slope, but that the relative average activity of the two neural populations is insensitive to properties other than the ITD (e.g., the sound at the source). Thus, the application of the coding metaphor to tuning curve experiments leads to confusion between parameter sensitivity and information about the corresponding property in a broader context. It could be argued that information in a broad context at least requires sensitivity, but this is also technically incorrect[3] (see, e.g., Zylberberg 2018).

2.3.2. Encoding visual stimuli

In visual neuroscience, theories of neural coding are based on heterogeneous tunings. There are several theories of population coding of stimulus properties in the visual cortex (Jazayeri and Movshon 2006; Pouget et al. 2003). One influential theory, the "Bayesian brain" hypothesis (Knill and Pouget 2004), postulates that neural activity represents the probability distribution of the stimulus property, which the brain can manipulate to perform statistical inference. A key assumption in this and other coding theories is that the firing rate of neurons is a context-free function of stimulus properties. This assumption appears explicitly in the models and in the way the brain is proposed to compute with those representations. For example, one variation of this theory proposes that the brain computes the log likelihood of a stimulus property by summing the activity of neurons weighted by the logarithm of each neuron's tuning curve (Jazayeri and Movshon 2006). This operation is described as a "simple neural readout strategy," because it involves only summation and multiplication by fixed weights. As already discussed, the problem is that in reality, tuning curves are defined for a specific experimental condition; they are not context free. Therefore, either the computation of the log likelihood will be systematically incorrect for all other conditions, or the weights used in the readout must be adapted to correspond to the tuning curve of each condition by an undescribed mechanism, in which case the readout cannot possibly be described as a "simple neural readout."

To what extent do tuning curves depend on context? As it turns out, to a large extent. It has long been known that properties of sensory neurons adapt to input statistics (Barlow et al. 1957; Hosoya et al. 2005). In the primary visual cortex, responses to local orientation depend on the surrounding context (Bolz and Gilbert 1986; Hubel and Wiesel 1968). Tuning properties of visual cortical neurons (not just the gain) depend on cognitive context, including the task the animal is doing (Gilbert and Li 2013), locomotion (Pakan et al. 2018), and prior presentation of sounds (Chanauria et al. 2018). Current evidence indicates that the activity of neurons is sensitive to stimulus properties (the technical sense of coding) but cannot be considered as context-free symbols that stand for the corresponding properties (the representational sense of coding). Can neural coding theories of perception accommodate for this fact? It would require that in every context, changes in encoding (stimulus-response properties) are mirrored exactly by changes in decoding (computations performed on neural activity, e.g., the "simple neural readout"). No mechanism has been proposed to achieve this (see also next section).

Theories of neural coding have the ambition to explain some aspects of perceptual behavior, namely, results of psychophysical experiments. Again, this requires that a link is made between the neural code and behavior. This link involves ideal observers; for each possible task there is an optimal way to decode neural activity into the variable of interest, which uses detailed elements of the experimental design. Critically, this link with behavior is not considered part of the model because it is assumed that it belongs to the reader of the neural codes.[4] Thus, the behavioral predictions of the coding theories critically rely on linking propositions whose validity or plausibility is not addressed. To be clear, the questionable assumption is not so much whether behavior or perception is optimal in some way (Rahnev and Denison 2018), but whether the activity of a neuron is something that is read and manipulated as if it were a register of a processor and not simply something that the neuron is doing at a particular time (acting on other neurons).

When the brain is engaged in solving a particular visual task, the activity of neurons depends specifically on object properties relevant to that task (Gilbert and Li 2013). This seems entirely logical if we see neurons as collaborating to solve a task. In contrast, it is surprising if we see the visual cortex as encoding the world and the remainder of the brain as dealing with this representation to guide actions. Thus, thinking in terms of coding seems to obscure rather than clarify understanding.

2.4. Can neurons encode variables?

It could be objected that the problem of contextual dependence of tuning curves calls only for a minor amendment to the mainstream neural coding theories, which is to consider that contextual variables are encoded too. This would require more complicated decoding schemes, but not fundamentally different theories.

For example, it could be proposed that populations of cones jointly encode wavelength and intensity, and both can be decoded from the joint activity of cones. But to decode cone activity into wavelength, it must be known that a monochromatic light is being presented. In natural experience, light is not monochromatic; it has a continuous spectrum, and the transduced current depends on the convolution of the spectrum of incident light with the absorption spectrum of the photoreceptor. In those cases, cones cannot possibly encode wavelength, even jointly, because there is no such thing as the wavelength of a patch of visual scene. Thus, the activity of cones is not sufficient to infer wavelength. A critical element of context that also needs to be encoded is the fact that a monochromatic light is being presented.

Similarly, a cat's neuron may encode the orientation of a bar only in conjunction with the information that a bar is being presented. That information does not take the form of a variable, but perhaps of a model of the experiment. But models are not variables; rather, they define variables. Thus, a perceptual scene cannot be represented by a set of variables, because this leaves out what defines variables. This missing aspect corresponds to object formation and scene analysis, two fundamental aspects of perception that are not addressed by coding: There is no object property to be encoded if there is no object.

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

Consider the Bayesian brain hypothesis: "the brain represents information probabilistically, by coding and computing with probability density functions" (Knill and Pouget 2004). This presupposes that there is a set of predefined variables to which probability is attributed – examples of variables are the position of an object and the orientation of a visual grating. If neurons encode variables, then what encodes the definitions of those variables, and what do neurons encode in situations where those variables are not defined? We can imagine that such theories might apply to the representation of eye position, for example, because the eye is always there and its position is always defined. This is not the case of objects of perception in general.

Similarly, influential models of working memory propose that memory items are stored and encoded in the persistent activity of neurons tuned to the underlying stimulus property, for example, the spatial position of an object (Constantinidis and Klingberg 2016). This provides a way to store graded properties, such as the position of a visual target or the pitch of a musical note. But suppose there is a neural network in my brain that is storing the number 100. What have I memorized? Clearly not the same piece of information if this number is the area of my apartment in square meters or the height of my son in centimeters. To store the information, one needs not only the number but also what it refers to. Can the persistent activity of tuned neurons store that information? It can if there is a network of neurons tuned to the area of my apartment and another tuned to the height of my son.

Perception and memory cannot just be about encoding stimulus properties because this omits the definitions of those properties and of the objects to which they are attached. But could it be that neurons encode more abstract "internal variables" that somehow describe the external world? Such is the claim of predictive coding (Rao and Ballard 1999) and related propositions such as the free energy principle (Clark 2013; Friston 2009). In these theories, neural coding is described as a statistical inference process, where neurons encode the inferred value of internal variables of a generative model of the inputs, for example, the retinal image. Technically, this essentially means that the code is a parametric description of the image (e.g., a Fourier transform). Described at this technical level, the theory seems to have little to say about perception or behavior. But the intended scope extends as these internal variables are described as the "causes" of the sensory input, and the process of encoding is referred to as "inferring the hidden causes." The sensory input is caused by things in the world, so an internal variable can be considered a cause only if it is assumed to encode properties of objects in the world as in the Bayesian brain hypothesis. Again, this is incoherent because no perceptual scene can be fully specified by the properties of its objects; one needs first to define objects and their properties, and these definitions are not conveyed by the variables. "Cause" must then be understood in the strict technical sense of variable of a statistical function, which has little to do with the usual sense of "cause." Thus, this use of the term *cause* appears to be another case of a metaphoric extension of the technical vocabulary. As Chomsky (1959, p. 30) observed, "This creates the illusion of a rigorous scientific theory with a very broad scope, although in fact the terms … [have] at most a vague similarity of meaning." We will return to predictive coding theory in the next section.

Neurons encode stimulus properties according to the technical sense of the metaphor. To acquire a broad scope, the metaphor drifts into the representational sense, according to which neurons convey information about the said properties to the rest of the brain. But neural activity can only be interpreted as properties once the interpretative framework is provided. Critically, this framework is not contained in the coding variables. In what sense do neural codes constitute information for the brain if their meaning lies outside the encoded messages and varies depending on situations? Where do ideal observers obtain the information necessary to decode the messages? In the next section, I argue that the coding metaphor conveys a very particular notion of information, which is information by reference, and that this is not the kind of information relevant to perception and behavior.

3. Do neural codes constitute information about the world?

3.1. Codes as information by reference

The coding metaphor assumes that neural codes represent information about the world, which the brain uses to produce adapted behavior. This sense is implied by the use of ideal observers in the neural coding literature and, more generally, by the presumption that the brain "decodes" neural responses or "extracts information" from them. In what sense is the neural code "information" about objective properties of the world? According to the technical sense of coding, it is information in the sense that these properties can be inferred from neural activity. Methodologically, this inference is made by the experimenter, who confronts these properties with measurements of neural activity. But by using the term *neural code* and by comparing the output of ideal observers with psychophysical measurements, we imply that the brain must also make this inference.

This raises the issue of "the view from inside the box" (Clark 2013): How is it possible for the nervous system to infer external properties from neural activity if all that it ever gets to observe is that activity? In fact, what does it even mean that a neural network infers external properties (e.g., the direction of a sound source), given that those properties do not belong to the domain of neural activity? This is related to the *symbol grounding problem* (Harnad 1990b): How do spikes, the symbols of the neural code, make sense for the organism?

A fundamental issue with the coding metaphor, as it applies to the brain, is that it conveys a very particular notion of information, information by reference; the meaning of the encoded message is that of the original message to which it refers. Shannon (1948, p. 379) made this very clear when he defined his mathematical notion of information:

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is, they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem.

But the semantic aspects are precisely what is relevant to the biological problem: How does the brain know what the codes refer to?

One possibility is that the meaning of neural codes is implicit in the structure of the brain that reads them; the brain understands neural codes because it has evolved to do so. There are at least two objections that make this proposition implausible. First, there is considerable plasticity, including developmental plasticity, both in the nervous system and in the body, which makes the idea of a fixed code implausible. An impressive example is the case of a patient born with a single brain hemisphere, who has normal vision in both hemifields, with a complete reorganization of brain

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

structure (Muckli et al. 2009). This plasticity implies that the "reader" of neural codes must learn their meaning, at least to some extent. Second, it might be imagined that the meaning of a neural code for eye position might be fixed by evolution – that there are fixed motor circuits that control the eye based on that fixed neural code. But how could this be true of a neural code for the memory item "my apartment is 100 meters square"?

To see that Shannon information cannot be the notion of information relevant to understanding perception, consider the following experiment of thought, which I call the paradox of efficient coding. Suppose that all information about the world (including efferent copies) is encoded by a set of neurons. From the heuristic that biological organisms tend to be efficient, we now postulate that neurons transform their inputs in such a way as to transmit the maximum amount of information about the world, in the sense of Shannon; this is the efficient coding hypothesis (Barlow 1961; Olshausen and Field 2004). This means that all redundancy is removed from the original signals. If it is done perfectly, then encoded messages are indistinguishable from random by the organism. Therefore, the perfectly efficient code cannot be understood by its reader.

It is indeed paradoxical that when we maximize the amount of information carried by code, we find something that provides no information at all to the reader. This is because the notion of information implied by the phrase "neurons encode information" is information by reference to the inputs, a kind of information that is accessible only to an external observer. This is not the right way to address the representational problems faced by the organism. As Bickhard has argued (Bickhard 2009; Bickhard and Terveen 1996), "encodingism" fails to provide an adequate notion of representation because it does not allow the possibility of system detectable error; there is no way for the system to know whether the representation is in error.

Again, the coding metaphor appears to promote a semantic drift, from the technical sense of information as defined by Shannon to a broader sense of information that might be useful for an organism. The neural coding metaphor is so prevalent in the neuroscience literature that the notion of information it carries seems to be the only possible one: "the abstract definition of information is well motivated, unique, and most certainly relevant to the brain" (Simoncelli 2003, p. 145). Next, I discuss alternative notions of information that are more relevant to the brain.

3.2. Information as subjective laws or internal models

How can there be any information about the world without direct access to the world? John Eccles, a prominent neurophysiologist, expressed the problem in the following terms (Eccles 1965, p. 322):

> In response to sensory stimulation, I experience a private perceptual world which must be regarded, neurophysiologically, as an interpretation of specific events in my brain. Hence I am confronted by the problem: how can these diverse cerebral patterns of activity give me valid pictures of the external world?

To him, the logical solution was a form of dualism, much like Cartesian dualism, except he did not believe that the interaction between mind and brain occurred at a single place (Descartes' pineal gland). Dualism is a natural solution if neural activity is thought to encode information by reference to the external world, because the external world belongs to a different domain.

A number of philosophers and psychologists have proposed alternative solutions. O'Regan and Noë (2001) proposed the analogy of the "villainous monster." Imagine you are exploring the sea with an underwater vessel. A villainous monster mixes all the cables; hence, all the sensors and actuators are now related to the external world in a new way. How can you know anything about the world? The only way is to analyze the structure of sensor data and their relationships to actions that you can perform. If dualism is rejected, then this is the kind of information that is available to the nervous system. A salient feature of this notion of information is that, in contrast to Shannon's information, it is defined as relations or logical propositions: If I do action *A*, then sensory property *B* occurs; if sensory property *A* occurs, then another property *B* occurs next; if I do action *A* in sensory context *B*, then C occurs.

Gibson (1979) previously developed a related psychological theory. While criticizing the information-processing view of perception, he argued that there is information about the world present in the *invariant structure* of sensory signals: "A great many properties of the [optical] array are lawfully or regularly variant with change of observation point, and this means that in each case a property defined by the law is invariant." Clearly, he did not mean information in the sense of communication theory, but rather in the sense of scientific knowledge. A set of observations and experiments provide information about the world in the form of laws that relate observables (sensory signals) between them and with possible actions. This form of information is intrinsic; I proposed calling this set of laws the subjective physics of the world (Brette 2016), which is related to von Uexküll's (1909) Umwelt. A related view, formalized by theoretical biologist Robert Rosen (1985), is that biological organisms build an internal model of the world in which the variables are sensory signals. This view addresses the symbol grounding problem by mapping sensory signals to elements of an internal model. The signals make sense in reference to that model; they are not mapped to externally defined properties. In Bickhard's (2009; 2015c) interactivist model, representations are based on anticipatory relations between internal processes.

Crucially, relations between observables are precisely what neural coding theory considers as redundancy, which ideally should be eliminated. In contrast, in the alternative view discussed here, relations constitute information. This point was made by Thompson (1968, p. 305). "It is our subjective habit to organize the individual elements of our experience, to cross-correlate these elements to others distant in space and time, and it is only after this process of imposing organization that we feel informed." The number 100 does not really constitute information; only when I have inserted it into my internal model of the world by saying that it is the area of my apartment in square meters does it become information.

3.3. Subjective physics of the Martian iguana

To make this point more concrete, I will discuss an example adapted from Brette (2016). Consider a fictional organism with two ears – let us call it a Martian iguana in reference to Dennett (1978) (Fig. 4A). The iguana is fixed on the ground, and there is another organism – let us call it a frog – which produces sounds. The frog is usually still and produces some random sounds repeatedly, but occasionally it jumps to a new position. The question is: To what kind of information can the iguana have access based on the acoustical signals at the two ears?

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

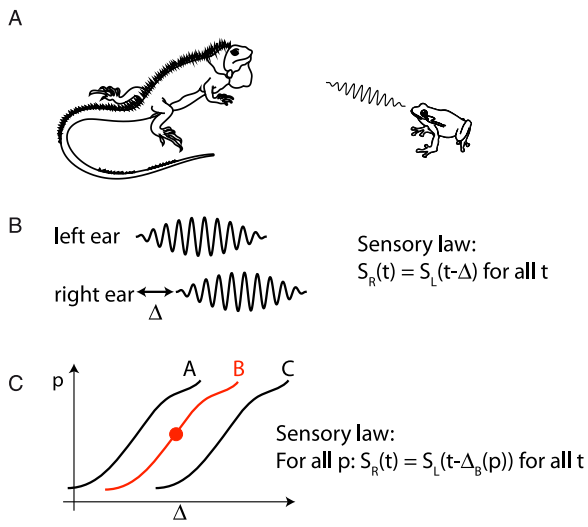


Figure 4. Subjective physics of a fictional iguana. (A) The (blind) iguana listens to sounds produced by the frog, which occasionally jumps to a new position. (B) When there is a sound, the iguana can notice that the acoustical signals at its two ears follow a particular law: $S_R(t) = S_L(t - \Delta)$ for all t. (C) If the iguana can move its head, it can also notice that the delay Δ changes in a lawful way with the proprioceptive signal p. This relation defines the frog's position for the iguana. When a sound is heard, the iguana can infer the frog's position; that is, it can infer how Δ would change if it were to move its head.

When a source produces a sound, two sound waves S_L and S_R arrive at the two ears, and these two sound waves have a particular property: they are delayed versions of each other ($S_L(t) = S_R(t - \Delta)$) (Fig. 4B). In Gibsonian terminology, there is "invariant structure" in the sensory flow, which is to say that the signals obey a particular law. Thus, the sensory world of the iguana is made of random pairs of signals that follow particular laws that the iguana can identify. This identification is what Gibson called the "pick-up of information." Evidently, "information" is not meant in the sense of Shannon but in the sense of laws or models of the sensory input. Note that the model in question is not a generative model as in predictive coding, but relations between observables, as in the models of physics.

A first interesting aspect of this alternative notion of information is that the topology of the world projects to the topology of sensory laws. By this, I mean that two different sounds produced by the frog at the same position will produce pairs of signals (S_L, S_R) that share the same property (the sensory law). This can be assessed without knowing what this property corresponds to in the world (i.e., the frog's position).

Thus, the iguana can observe sensory laws that have some particular properties, but do these laws convey any information about where the frog is? For an external observer, they certainly do, because the delay Δ is lawfully related to the frog's position. For the iguana, however, they do not because that lawful relation cannot be inferred from simply observing the acoustical signals. Therefore, this organism cannot have any sense of space, even though neural coding theories would pretend that it does, based on the correspondence between frog position and activity of the iguana's auditory neurons.

Let us now consider in addition that the iguana can turn its head (Fig. 4C). It can then observe a lawful relation between a proprioceptive signal (related to the head's position) and the observed delay Δ, which holds for some time (until the frog jumps to another position). Now when the iguana observes sounds with a particular delay, it can infer that if it were to move its head, then the delay would change in a particular predictable way. For the iguana, the relation between acoustical delay and proprioception *defines* the spatial position of the frog. We note that the perceptual inference involved here does not refer to a property in the external world (frog position), but to manipulations of an internal sensorimotor model.

Thus, the kind of information available to an organism is not Shannon information (correspondence to external properties of the world), but internal sensorimotor models. The interest of such models for the animal is that they can be manipulated to predict the effect of hypothetical actions.

3.4. Predictive coding and generative models

Predictive coding theory and its derivatives (Clark 2013; Friston 2010; Rao and Ballard 1999) propose that the brain encodes an internal model, which predicts the sensory inputs.[5] This seems to resemble the proposition put forward in the previous section. More precisely, neurons are thought to encode the variables of a hierarchical model of the inputs, in which higher-order neurons encode their prediction of the activity of neurons lower in the hierarchy, down to the sensory inputs. This prediction is subtracted from the input of lower-order neurons, so only the prediction error remains. This leads to a compressed representation of the inputs, and in this sense, it is a type of efficient coding theory.

This particular kind of model is called a generative model because its maps internal variables to the observables (sensory inputs), in contrast to the models of physics, which take the form of relations between observables (e.g., the ideal gas law, $PV = nRT$). Generative models are not the kind of internal models described in the previous section.

Consider the iguana with a fixed head. A generative model of the sensory inputs would map two internal variables S (sound) and Δ (interaural delay) to the two acoustical inputs S_L and S_R, as $S_L(t) = S(t)$, $S_R(t) = S(t - \Delta)$. Neural activity encodes not the model itself but the coding variables Δ and S. In particular, neurons encode the entire sound S, even though it carries no information for the iguana (S is, by construction, random). This appears to contradict the claims of predicting coding theory: "To successfully represent the world in perception … depends crucially upon cancelling out sensory prediction error" (Clark 2013, p. 7). Indeed, the success of a predictive code is evaluated by its ability to represent the input in a pictorial sense (as if it were a painting), but in this example, the numerical value of the signals provides no useful information beyond the relations they obey.

Consider now the case where the iguana can move its head. The internal model discussed in the previous section is $S_R(t) = S_L(t - \Delta_x(p))$ for all t, where x is the frog's position (Fig. 4C). The usefulness of this model stems from the fact that it can be manipulated; that is, on hearing a sound, the iguana can infer that, if it were to move its head to a new position p, the relation obeyed by the auditory signals would change in a predictable way. For example, the iguana can move its head so that $S_R = S_L$ ("the frog is in front"). Thus, the kind of prediction that this model can produce is about relations between signals and not about the numerical value of the signals.

On the other hand, a generative model would map the coding variables S, x, and p to the sensory inputs $S_L(t) = S(t)$ and $S_R(t) = S(t - \Delta_x(p))$. This mapping is referred to as *prediction* and is instantiated by the feedback from higher-order neurons to

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

lower-order neurons. This is not the same as predicting what action would make the two signals S_L and S_R match, which brings us to a discussion of the term *predictive* in predictive coding. The appeal of predictive processing is that making predictions seems to be a prerequisite to goal-directed behavior and, thus, a fundamental aspect of behavior. In fact, several authors have argued that anticipation is not simply a property of nervous systems, but even a fundamental property of life (Maturana and Varela 1973; Rosen 1985). For example, the iguana can predict how some properties of its sensory inputs should change if it were to turn its head. Consider this other example in human behavior: When someone is facing a cliff, she tends to slightly lean backward, because this posture makes it easier to move backward if necessary (Le Mouel and Brette 2017). But this is not at all the technical sense of "prediction" in predictive coding, as argued by Anderson and Chemero (2013). A neuron "predicts" the sensory inputs in the sense that its firing correlates with them; more specifically, a spike produced by a neuron leads to a subtraction of the expected input of a target neuron, which is the input occurring now, or possibly if we incorporate conduction delays, what will be happening after a fixed delay. This is not the kind of prediction implied by an anticipatory postural adjustment: If I change my posture in this way, then it will be easier to move backward in the hypothetical event that my balance is challenged.

In fact, what is useful for the organism is not literally to predict what will happen next, but rather what *might* happen next, conditionally on the actions I can do, so that I can select the appropriate action. But this requires manipulation of the model. For example, selection of an action requires instantiating the internal model with several possible values of action and then calculating the expected sensory variables. But this contradicts the proposition that neurons encode the "causes" of current sensory signals; to manipulate the model, encoding neurons would then have to be somehow disconnected from the sensory stream.

Technical work on predictive coding has focused exclusively on the technical senses of prediction and coding (correspondence); therefore, there is no empirical evidence that such codes might allow the organism to form predictions in a broader sense, nor is there any indication of how a theory based on neural coding might in principle explain anticipatory behavior.

3.5. Can neural codes represent structure?

The kind of representation of the world useful for adapted behavior is a structured internal model. Can neural codes possibly represent that structure? Memories and percepts are thought to be encoded by cell assemblies. In its basic and most popular form, a cell assembly is simply a specific subset of all neurons. When neurons of a cell assembly activate, the corresponding percept is formed (possibly indirectly by the activation of target neurons). This is the basic assumption of associative neural models of memory (Tonegawa et al. 2015): Retrieving a memory consists of triggering activity in part of the memory-specific cell assembly (or "engram cells"), which then leads to the activation of all neurons in the assembly.

One problem with cell assemblies, in this simple form, is that they are unstructured and, thus, cannot represent structured internal models. The cell assembly model is analogous to the "bag of words" model in text retrieval, where a text is represented by its set of words and all syntax is discarded. In essence, a cell assembly is a "bag of neurons." This causes a problem in representing not only the lawful structure of the world, but also the structure of any given perceptual scene. Consider, for example, the simple visual scene depicted in Figure 5. There is Paul, a person I know, wearing a new shirt, driving a car (Fig. 5). What is important here is that a scene is not simply a "bag of objects"; objects have relationships with each other, and there are many possible different relationships. For example, there is a car and there is Paul, and Paul is in specific relationships with the car, both a physical relationship (a particular posture within the car) and a functional relationship (driving it). Some of my behavior depends on identifying these relations because, for example, I can talk about them, and so if behavior relies on neural codes, then those codes should represent relations, not just the pixels of the image.

But cell assemblies cannot represent these relations. Suppose there is a cell assembly that encodes "Paul" and another one that encodes "car." To encode the driving relation between Paul and the car, a cell assembly that encodes "driving" would also be needed, but that assembly should also somehow refer to the two assemblies representing Paul and the car, and this is something that cannot be done with an unstructured bag of neurons (mathematically, one would need a labeled graph and not simply a subset of nodes).

This is related to the "binding problem," although it is broader. If it is true that any given object is represented by the firing of a given assembly of neurons, then several objects should be represented by the firing of a larger assembly of neurons, the union of all assemblies, one for each object. Several authors have noted that this may lead to the "superposition catastrophe" (von der Malsburg 1999); that is, there may be different sets of objects whose representations are fused into the same big assembly. One proposition is that the binding problem could be solved using retinotopic position as an object label; that is, neurons do not encode features but encode the conjunction of feature and retinotopic position (Kawato 1997). However, this objection does not address the broader point, which is that cell assemblies encode objects or features to be related, but not the relations between them. In fact, it is known that current connectionist models,

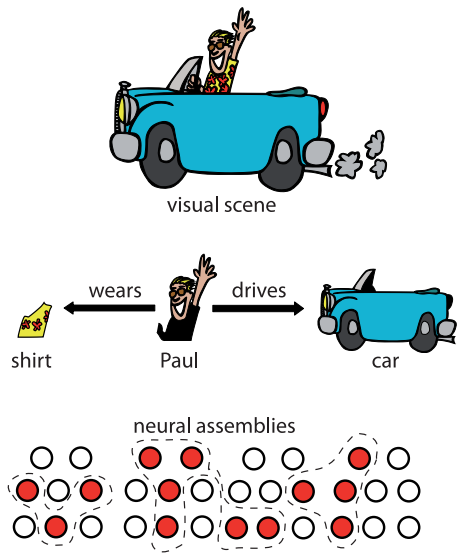


Figure 5. Perceptual scenes are highly structured. For example, there is Paul (person I know), driving a car and wearing a new shirt. Representing this scene by the firing of neural assemblies raises two issues: (1) It may be difficult to split active neurons into the correct assemblies (superposition catastrophe) and, more importantly, (2) the structure of the scene (relations shown by *arrows*) cannot be represented in this way.

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

which are designed to optimally implement the idea that features are represented by the activity of one or several cells, cannot be trained to detect very simple relations between shapes in an image (Ricci et al. 2018).

The binding problem has led several authors to postulate that synchrony is used to bind the features of an object represented by neural firing[6] (Singer 1999; von der Malsburg 1999). This avoids the superposition catastrophe because at a given time, only one object is represented by neural firing. Synchrony is indeed a relation between neurons (mathematically, an equivalence relation). There are a few other examples in the neuroscience literature where synchrony is used to represent relations, although they are not usually cast in this way. One is the Jeffress (1948) model of ITD coding (Fig. 6A). In that model, neurons receive inputs from monaural neurons on the two sides, with different conduction delays. When input spikes arrive simultaneously, the neuron spikes. Thus, the neuron spikes when the two acoustical signals at the two ears are such that $S_L(t) = S_R(t - d)$, where d is the conduction delay mismatch between the two ears. Physically, this corresponds to a sound source placed at a position such that it produces an ITD equal to d. In this model, the neuron's firing indicates whether signals satisfy a particular sensory law.

This interpretation of the model has been generalized with the concept of "synchrony receptive field" (Brette 2012), which is the set of stimuli that elicit synchronous responses in a given group of neurons (Fig. 6B). One considers two neurons A and B that convert their time-varying inputs into precisely timed spike trains, where their inputs are seen as transformed versions $N_A(S)$ and $N_B(S)$ of the stimulus S (N_A and N_B are fixed and correspond to the receptive fields of the neurons). Synchrony between A and B then reflects ("encodes") the sensory law $N_A(S) = N_B(S)$. This framework has been applied to pitch perception (Laudanski et al. 2014) and to sound localization in realistic environments (Benichoux et al. 2015; Goodman and Brette 2010).

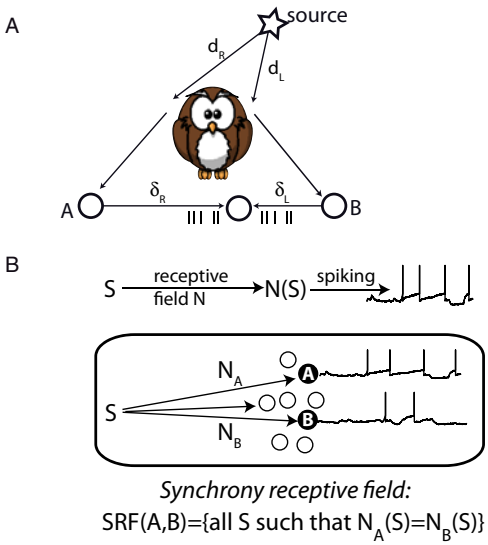


Figure 6. Neural representation of structure (adapted from Brette 2012). (A) Jeffress' model of sound localization. The sound arrives at the two ears with delays d_L and d_R. It is then transduced into spike trains that arrive at a binaural neuron with delays d_L and d_R. Synchrony occurs when $d_R - d_L = \delta_L - \delta_R$, making the neuron fire. (B) Synchrony receptive field. The response of a neuron to a stimulus is described as filtering of the sensory signal S through the receptive field N, followed by spiking. The synchrony receptive field of two neurons A and B with different receptive fields N_A and N_B is defined as the set of stimuli that elicit synchronous responses in these neurons.

Although synchrony can represent relations, neither binding by synchrony nor synchrony receptive fields solve the general problem (even theoretically), because only one type of relation can be represented by synchrony, and a symmetrical one: Does Paul drive the car, or does the car run over Paul? The fact that sentences can represent relations motivates the idea that the temporal structure of neural activity (e.g., the sequence of activated neurons, much like a sequence of words) could perhaps provide the adequate basis for structured neural representations (Buzsáki 2010). But this possibility remains speculative, and in particular, it remains to be demonstrated whether such hypothetical structures have the quality of representations that the brain can manipulate.

4. The causal structure of the coding metaphor

In the previous sections, I have argued that neural coding theories generally rely on the representational sense of the metaphor, the idea that neural codes are symbols standing for properties that the brain manipulates, but no evidence has been provided that this sense is valid. Worse, there is empirical evidence and theoretical arguments to the contrary.

Here I focus on a deeper problem with the neural coding metaphor. A striking characteristic of this metaphor is that it is a way to think about the brain independently of its causal structure. When we say, for example, that neurons encode the location of sounds, we talk about the activity of neurons without making any reference to the result of that activity or to the system of which the neurons are a component. I now examine the implications of this fact.

4.1. The dualistic structure of the coding metaphor

The coding metaphor has a dualistic structure. It structures the function of the brain into two distinct and dual components: the component that encodes the world into the activity of neurons, and the dual component that decodes that activity into the world or into actions in the world, as illustrated by the following examples: "Information that has been coded must at some point be decoded also; one suspects, then, that somewhere within the nervous system there is another interface, or boundary, but not necessarily a geometrical surface, where 'code' becomes 'image'" (Somjen 1972, p. 3). "[I]nterpretation of the encoded information, typically consisting of its recoding by a higher-order set of neurons or of its "decoding" by an effector" (Perkel and Bullock 1968, p. 307). "A stimulus activates a population of neurons in various areas of the brain. To guide behavior, the brain must correctly decode this population response and extract the sensory information as reliably as possible" (Jazayeri and Movshon 2006, p. 690). "[T]he brain typically makes decisions … by evaluating the activity of large neuronal populations" (Quian Quiroga and Panzeri 2009, p. 173). "Ideal observers" used in many studies implement this dual-decoding brain.

Using the coding metaphor does not necessarily mean believing in dualism of body and mind,[7] but its dualistic structure has important consequences when it comes to understanding function. The two dual components (encoding/decoding) are indistinguishable in behavior, because no behavior involves just one of them. How then is it possible to attribute function to neural codes? How is it possible to draw conclusions about the neural

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

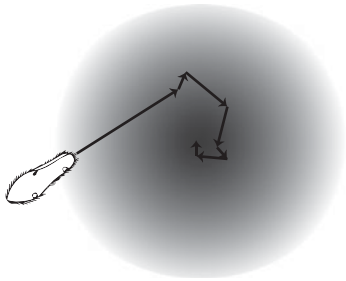


Figure 7. Spatial cognition in *Paramecium*, a "swimming neuron." *Paramecium* finds a chemical source by switching to a new random direction when concentration decreases.

basis of behavior from properties of neural codes, independently of the system in which the neurons are embedded? This is only possible by making an additional assumption, namely, that the encoding component has a function by itself (representing the inputs), somehow assigning the status of organ to a part of the nervous system. But there is no indication that the brain can be functionally decoupled in this way; neuroanatomy rather seems to invalidate this hypothesis.

To illustrate this point, I now discuss a concrete biological example. *Paramecium* is a unicellular organism that swims in stagnant fresh water using cilia and feeds on bacteria. It uses different kinds of sensory signals, including mechanical signals to avoid obstacles and chemical signals to localize food (Jennings 1906). To a first approximation, it alternates between straight courses and sudden random changes in direction (Fig. 7). It turns out that each change in direction is triggered by a spike produced by voltage-gated calcium channels (Eckert 1972). To find a chemical source, *Paramecium* uses a simple method: when concentration decreases, the membrane is depolarized by chemical receptors and a spike is produced (with some stochasticity), triggering a change of direction (similar to chemotaxis in *Escherichia coli*). This is of course a simplified description of *Paramecium* physiology and behavior, but for the sake of argument, we will consider an organism that functions in this simple way.

Paramecium is thus a sort of swimming neuron. Spiking activity varies lawfully with sensory signals (concentration) and, hence, encodes them in the same sense as a visual cortical neuron encodes visual signals. As for sensory neurons of the brain, we may argue that if the organism can navigate efficiently in its environment, then the spikes must contain information about that environment. Thus, it seems that the coding metaphor applies equally well to this swimming neuron as to any typical case in neuroscience.

Let us now think about functional questions. As an organism, *Paramecium* may have goals, for example, finding food. We may hypothesize that it achieves this goal efficiently, for example, by finding food as quickly as possible. To this end, sensory signals must be transformed into spikes in a specific way, which depends both on the goal (to move toward or away from a source, to look for food or to sleep, or to look for a mate) and on the effect of spikes on the organism's actions. Thus, there is a way to organize this system so that it achieves its function appropriately, which determines the transformation of inputs into spikes, that is, the neural code.

But if we now think of the neural code independently of the organism and environment that host it, we draw different conclusions. If the function of this neuron is to encode its input, then we may hypothesize that it achieves this function efficiently. This prescription determines a neural code that is specified by the statistics of inputs. Here the code depends neither on the goals of the animals nor on the effect of spikes on the organism's actions. It follows that this efficient code does not match, in general, the neural code that is adapted for the organism's goal. This mismatch occurs because function can be meaningfully ascribed to the organism as a system, but not necessarily to the components of this system.

This sensorimotor system is arguably much simpler than the brain; nevertheless, it demonstrates that the function of neurons cannot be meaningfully framed in terms of coding just because they respond to sensory stimuli. There is no indication that the brain is special in that it can be meaningfully separated into two dual components with independent functionality.

4.2. Coding versus causing

The *Paramecium* example highlights the fact that the neural coding metaphor is a way to think about the brain that is disconnected from its causal structure. Yet by postulating that neural codes are representations, we imply that these codes have a causal impact on the brain. This is also the case when neural codes are considered simply as transformations of inputs rather than explicit representations, as in Perkel and Bullock (1968, p. 227): "The problem of neural coding is defined as that of elucidating the transformations of information in the nervous system, from receptors through internuncials to motor neurons to effectors." But does coding imply causing?

Consider, for example, the BOLD (blood oxygen level-dependent) signal, a property of blood used for functional brain imaging because it covaries with neural activity. The signal encodes visual signals in the same technical sense that the firing of neurons encodes visual signals. For example, one can "decode" the image from this signal (Naselaris et al. 2009). Yet, visual perception is not caused by the BOLD signal, which is why we do not consider that it is an internal representation used by the brain. Thus, not all coding variables have causal powers.

Consider the firing rate-versus-spike timing debate (Brette 2015; Kumar et al. 2010). This debate is generally formulated as follows: "Does the brain use a firing rate code or a spike timing code?" As the previous example illustrates, this is a largely irrelevant question because it focuses on correlations between stimuli and observables. We may as well ask Does the brain use the BOLD code? The relevant question is rather whether those observables have a causal role in the activity of the brain, and this involves a different set of arguments and answers (see Brette [2015] for a discussion). To see why, consider a sensorimotor system whose function is well understood in relation to its electrical activity: the heart (Fig. 8). The heart operates like a pump to circulate blood in two phases: the two atria contract, pushing blood into the ventricles (diastole); then the two ventricles contract, pushing blood into the pulmonary arteries (systole). These contractions are triggered by excitable cells in the atria and ventricles. For the heart to operate as a pump, cells in the two atria must spike synchronously, but out of phase with cells in the two ventricles. But the heart also responds to sensory stimulation. For example, the heart beats at a faster pace when we run. This means that the excitable cells of the heart encode running speed in their firing rate, in the technical sense. If we now look at the coding properties of these cells, we find that (1) firing rate is sensitive to running speed, (2) cells fire regularly, (3)

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

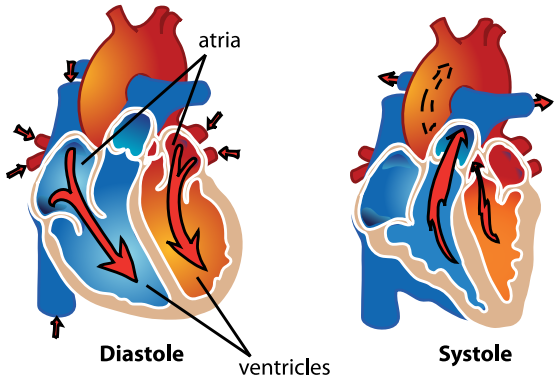


Figure 8. Operation of the heart. Atria simultaneously contract, triggered by synchronous firing of excitable cells; then ventricles simultaneously contract, pushing blood into the lungs.

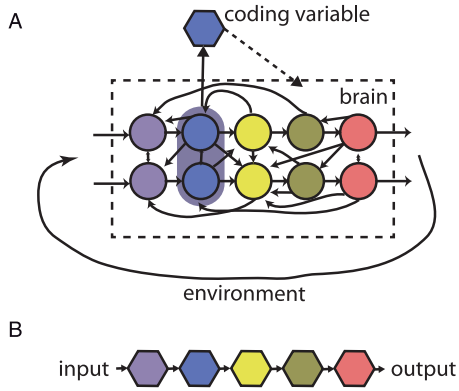


Figure 9. Causal structure of brain and neural codes. (A) The brain is a distributed dynamical system made of interacting neurons and is coupled to the environment by circular causality. A coding variable is a property of neural activity, which is implicitly assumed to have a causal effect on the brain. (B) Neural codes are linked together and with the world by linear causality.

spike timing is not reproducible between trials, and (4) spike timing (absolute or relative) carries no information about the stimulus beyond the rate. Thus, we would conclude that the heart uses a rate code. Yet, the temporal coordination of spikes is critical in this system; in fact, it is life critical. This paradox arises because the neural coding metaphor totally neglects the causal effect of spikes.

If we want to describe the operation of the brain in terms of neural coding, the relevant question is whether the causal structure of neural codes is congruent with the causal structure of the brain.

4.3. Causal powers of coding variables

The causal structure of the brain is sketched in Figure 9A. At a coarse description level, the brain is a dynamical system coupled to the environment by circular causality. At a finer description level, the brain is itself made of neurons, which are themselves dynamical systems coupled together. To a first approximation, the coupling is mediated by spikes, which are timed events.

Consider the proposition "the firing rate of neuron A encodes the location of a sound source," corresponding to some empirically observed correlation. The implication that this information is decoded by the brain relies on the presupposition that the coding variable "firing rate of neuron A" causally influences the future activity of the brain. Spikes, of course, have causal effects on the brain. But a neural coding variable ("firing rate of neuron A" or "relative activity of two neural populations") is a particular measurement of spiking activity, and the question is whether that particular measurement has causal powers.

Empirically, a coding variable is an aggregate variable based on measurements of spiking activity over some time, space, and possibly trials. An example of integrating over trials (and time) is a neuron that responds specifically to pictures of Jennifer Anniston in various poses (Quian Quiroga et al. 2005). But only on average: the coding variable is the median number of spikes across trials between 300 and 1,000 ms after stimulus onset. On a given trial, the neuron might not be firing at all. Unless the subject was not perceiving the actress in those trials, this implies that this neuron cannot encode the percept "Jennifer Anniston" in the sense of causing the percept. Rather, its firing correlates (on average) with the presentation of pictures of Jennifer Anniston, which is already a notable fact. Perceptual representations cannot be based on averages; percepts are experienced now, not on average. Neural codes based on averaging over trials do not have causal powers (see also Gomez-Marin and Mainen 2016). In the same way, a firing probability (one abstract way to define a neuron's firing rate) does not have causal powers[8]; only the occurrence of firing does.

An example of integrating over space (and time) is when we propose that the position of a sound source is encoded by the difference in total activity between the two symmetrical inferior colliculi (Grothe et al. 2010). This coding variable indeed varies when source position is changed (Thompson et al. 2006). Does it mean that it has causal powers, that is, that it determines sound localization behavior? It seems implausible, first, as previously discussed because it also varies with other properties of sounds, and second, because electrical stimulation in the inferior colliculus triggers orienting responses that vary with the place of stimulation, whereas stronger stimulation results in orienting responses that engage a larger part of the body (one pinna, both pinnae, and eyes, in order of recruitment) (Syka and Straschill 1970). Thus, there is no guarantee that a coding variable obtained by integrating over neurons has causal powers.

But the key difficulty is time. The course of a dynamical system is determined by its current state, which is characterized by state variables such as membrane potential and the state of ionic channels. Spikes, on the other hand, are events (something happening to the system) and not properties (some characteristic of the system). Therefore, spiking activity is not something defined at any point in time, which could give it the causal role of a state variable, but something that is measured over some predefined period (in the first example, 300 to 1,000 ms after stimulus onset). Empirically, a neural coding variable is necessarily anchored to the temporality of the experiment (some window of time after the onset of the stimulus). Once we have anchored variables in time, all possibility of physical interaction between coding variables disappears: if variables X and Y are defined over two different time windows T_1 and T_2, then there cannot be causal influence in both directions ($X \rightarrow Y$ and $Y \rightarrow X$). But neurons mutually influence themselves over timescales of a few milliseconds, without waiting for the coding variable to be defined or for a stimulus to be presented. Neural codes abstract time away, but temporality is critical to the operation of a dynamical system.

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

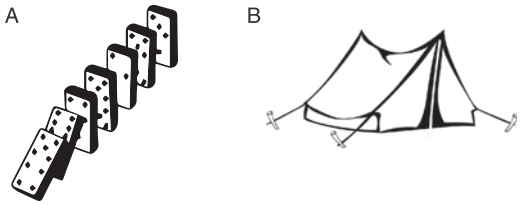


Figure 10. The causal structure of the neural coding metaphor is that of dominoes (A), but the causal structure of the brain rather resembles that of a tent (B).

4.4. Causal structure of neural codes

If neural coding variables are anchored in time, then the only possible causal structure linking coding variables is a linear sequence of transformations (Fig. 9B). This is implicit in Perkel and Bullock's (1968) definition of neural coding ("transformations … from receptors through internuncials to motor neurons to effectors"). This linear structure is also implicit in any claim that a stimulus is encoded, then decoded: cognition follows a linear causal flow,[9] from stimulus to perception to action (Hurley 2001). In such descriptions, the temporality of the physical system has disappeared and has been replaced by the discrete temporality of an algorithm, which is disconnected from physical time. In other words, this is an algorithmic description. But as van Gelder (1995) pointed out, dynamical systems cannot in general be mapped to algorithmic descriptions.[10]

The coding metaphor tries to match the causal structure of dominoes to the causal structure of a tent, where the states of different elements are co-determined (Fig. 10). In addition to the coupling of neurons, the brain itself is coupled to its environment; that is, there is circular and not linear causality (Fig. 9A). As Dewey (1896, p. 363) pointed out more than a century ago, "the motor response determines the stimulus, just as truly as sensory stimulus determines the movement." Many other authors in biology, psychology, philosophy, and robotics have argued that perception is not a one-way process but an interaction with the environment (Ahissar and Assa 2016; Brooks 1991a; Gibson 1979; O'Regan and Noë 2001; Powers 1973a). This makes the proposition that neural activity encodes stimuli questionable. In fact, it makes the very notion of stimulus questionable, as it seems to give no role for spontaneous activity other than noise (Deco et al. 2011), when autonomous activity is central to the organization of behavior. As a theory of cognition, the neural coding metaphor seems to embrace the most basic form of behaviorism.[11]

5. Conclusion

5.1. Summary

When I say that the heart is a pump, I propose a function for the heart (to circulate blood) and mechanisms by which blood is circulated; I propose specific ways in which elements of the heart interact by identification with elements of a pump. In effect, the pump is a model of the heart. A metaphor is not simply words arbitrarily chosen to designate an object; it is a model of the object (Lakoff and Johnson 1980a), and as such it deserves scrutiny as does any other model in science. Is coding a good model of brain function?

There are three aspects of the coding metaphor: correspondence, representation, and causality. Technical results are based on the first aspect, but their interpretation and claimed significance draw on the two other aspects, which are not subject to the same scrutiny. Many neural coding theories rely on the idea that the brain manipulates neural representations of stimulus properties, as if the variable of a neural code were a processor register that the brain can store, retrieve, and combine arbitrarily, while knowing what the variable refers to. But what is the evidence that such neural representations exist, and what is the evidence that the brain can manipulate spikes in this way?

Technically, it is found that the activity of many neurons varies with stimulus parameter, but also with sensory, behavioral, and cognitive context; neurons are also active in the absence of any particular stimulus. A tight correspondence between stimulus property and neural activity only exists within a highly constrained experimental situation. Thus, neural codes have much less representational power than generally claimed or implied. Behavioral significance is obtained only by making an implicit "linking proposition" (Teller 1984) that relates coding variables and behavior, which takes the form of a "decoder." The decoder, often an "ideal observer," is a hypothetical abstract construct whose biological basis is unspecified and whose existence is unquestioned, even though the decoder must incorporate key contextual aspects, including methodological details of the experiment, which defines the coding variables. Critically, the contextual dependence of neural codes cannot be solved by incorporating contextual variables in a broader neural code, because context is precisely what defines the variables. A perceptual scene cannot be fully defined as a vector of properties. Properties of what?

The notion of information implied by the coding metaphor is inappropriate in understanding perception and behavior, because it is information by reference to external symbols (Bickhard 2009). A more appropriate notion is information as organization (Thompson 1968), namely, relations between sensory signals and actions, forming a structured internal model. The relation between such structured models of the world and neural activity is unclear, but what is clear is that no neural coding theory proposed so far seems adequate, even in principle.

Ultimately, the neural coding metaphor is a way to think about the brain that is disconnected from its causal structure. The brain is a dynamical system coupled to the environment and is itself composed of coupled dynamical systems (neurons), whose interaction is mediated by spikes, which are timed events. The dualistic structure of the metaphor cuts through this organization and decides that one part of the brain can be understood independently of the way it interacts with the rest of the brain and independently of the way the brain interacts with the world. More fundamentally, a causal role is attributed to coding variables, but this is incoherent because coding variables are extended measurements of activity linked to the temporality of experiments; they are not causal variables of the underlying dynamical system. In conclusion, the causal structure of neural coding metaphor is incongruent with the causal structure of the brain. If neural codes have no causal power, then they cannot form a valid basis of a theory of brain function.

5.2. What else, if not coding?

Since the coding metaphor is so ingrained in neuroscience, how could it be possible to abandon it? What could it be replaced with?

First, that there is no simple substitute for neural coding does not make it a viable option. The neural coding metaphor is attractive partly because it resonates with Cartesian philosophy, as pointed out by Cisek (1999), and partly because it seems to fit with the computational view of the mind, the idea that cognition is the manipulation of symbols that represent properties of objects in the world. But the symbols provided by neural codes are not context free;

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

they are unstructured and they have no causal powers. They do not have the quality required by the computational view. Thus, the appeal of the neural coding metaphor is illusory. Even if it were possible to map brain activity to computational descriptions, neural codes would not provide the adequate mapping.

Similar arguments have been made against the idea of a genetic code (or genetic program), and the alternative route is to adopt a systemic approach (Noble 2008). The brain is a system, or, more accurately, the brain, body and environment are a system. This approach is precisely what the coding metaphor forbids, because it cuts through the system and uncouples its different components. Since it is a dynamical system, this view is related to the dynamical view of cognition (van Gelder 1998). But the specific point here is not so much that cognition is dynamic, but rather that its neural basis is a dynamical system and must be understood as such. It is a special kind of dynamical system in that it is composed of units (neurons), which are also dynamical systems. The causal role of spikes in this system is to mediate coupling between these dynamic units. They are transient events that are better understood as actions than as representations. A useful analogy then might be collective behavior: social insects are also dynamical systems coupled to each other by actions, and the collective behavior they display can be understood in terms of self-organization without resorting to the concept of coding (Bonabeau et al. 1997). This view should not be mistaken for an argument against representations in general, but more precisely against the classic view of representations as encodings. Bickhard (2015c), in particular, has made a case for representations as a form of normativity realized by anticipatory properties of internal processes.

In terms of neural modeling, this requires considering sensorimotor systems. The necessity of this level of analysis has been stressed by a number of authors who have developed alternative views on cognition (Ahissar and Assa 2016; Bickhard and Terveen 1996; Brooks 1991a; Gibson 1979; Hurley 2001; Maturana and Varela 1973; O'Regan and Noë 2001; Pezzulo and Cisek 2016; Powers 1973a). Paradoxically, it is customary in systems neuroscience to model perceptual abilities by considering only the corresponding sensory areas. We speak, for example, of the visual system as a set of anatomical structures from the eye to the visual cortex. But the visual system defined in this way is not actually a system if it is disconnected from the elements without which it cannot have any function. It follows that models of perceptual systems are in effect not biological models, but chimeras obtained by attaching a neural model of a sensory area to an abstract construct ("decoder") that maps the activity of neurons to descriptors of behavior and, often, to an even more problematic abstract construct ("encoder") that maps stimulus parameters to model inputs. This methodology embraces both behaviorism (neural activity is only responses to stimuli) and dualism (something else makes sense of neural activity). Instead, I suggest developing models of the full sensorimotor loop, "models that behave" (Gomez-Marin 2017). For example, instead of looking for neural codes of sound location, one could look for neural models of auditory orientation reflexes. Measurements of neural activity in stimulus-response experiments can be used to constrain and test such models, but they do not need to be the output of the model, nor do they need to be a causal variable in the model. To be clear, the issue is not about the amount of detail that needs to be incorporated. Models can be simplified or idealized, as any model needs to be. The issue is to respect the causal structure of brain and behavior and to see neural activity as what it really is: activity. Action potentials are potentials that produce actions; they are not hieroglyphs to be deciphered.

Notes

1. A strong correlation (or mutual information) between wavelength and current observed in the first case (Fig. 2B) may transfer to a negligible correlation in the second case (Fig. 2C) (Brette 2010).
2. A similar number of neurons are also tuned to small ITDs, especially in larger mammals such as cats (Goodman et al. 2013) (Fig. 1a).
3. Suppose, for example, that we observe the activity A of a neuron whose firing rate varies with parameter X as $A = X + Z$, where Z is an uncontrolled variable. If Z has large variance, A might hardly be correlated with X. But if we simultaneously observe $B = Z$, then we can recover X exactly ($B - A$), even though B is not correlated at all with X. There is no direct relation between parameter sensitivity assessed with a tuning curve and information in a broader context.
4. For example, Jazayeri and Movshon (2006) argue that the representation of the probability distribution of stimulus property (rather than simply the most likely property) allows the same code to perform different tasks, and they comment: "In contrast, previous models of sensory decoding were for the most part designed to account for a particular task" (p. 695). But the new proposition still requires a specific decoder for each task.
5. Not to be confused with predictive information, which is the mutual (Shannon) information between the past and future of a signal (Bialek et al. 2001; Palmer et al. 2015).
6. This proposition remains controversial because it is unclear whether synchronous firing across distant brain areas has causal powers (Merker 2013a) (see next part).
7. Nonetheless, the resemblance to Cartesian dualism is hard to miss. Indeed, Cisek (1999) argues that this dualistic structure has been inherited from Cartesian dualism, specifically that computationalism has replaced the nonphysical mind with a mechanistic cognition, while keeping the architecture unchanged (perception–cognition–action).
8. Except if the law of large numbers is used. This requires a number of assumptions (see Brette 2015).
9. Despite the explicit incorporation of feedback, hierarchical predictive coding still adheres to this general scheme, where stimulus is transformed into coding variables, which are then presumably used by some other process.
10. A notable exception being, of course, a computer executing an algorithm.
11. A variation that Gomez-Marin (2017) calls "neuralism."

Open Peer Commentary

Prediction, embodiment, and representation

István Aranyosi

Department of Philosophy, Bilkent University, 06800 Bilkent, Ankara, Turkey.
aranyosi@bilkent.edu.tr istvanaranyosi.net

doi:10.1017/S0140525X19001274, e216

Abstract

First, I argue that there is no agreement within non-classical cognitive science as to whether one should eliminate representations, hence, it is not clear that Brette's appeal to it is going to solve the problems with coding. Second, I argue that Brette's criticism of predictive coding as being intellectualistic is not justified, as predictive coding is compatible with embodied cognition.

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

Among the shortcomings that the metaphor of coding involves, Brette mentions its inability to truly function as a representation. At the same time, he seeks an alternative to coding in non-classical cognitive science, such as dynamic systems, ecological psychology, and embodied cognition, which, in their most radical and most interesting versions are precisely *anti*-representationalist approaches. How is the former complaint to be squared with the latter alleged solution? Brette does not tell us, but his critical discussion of predictive coding indicates that, ultimately, his problem with coding is the alleged intellectualism involved in it, hence, it is the alternative, embodied and embedded cognition theory that he thinks should be understood as currently the best remedy. He appears to think that an approach like predictive coding suffers from the same problems of intellectualism and inadequacy when it comes to how an organism perceives.

There are two problems with this view. One is that the embodied and embedded approach to the mind lacks anything close enough to agreement when it comes to whether representation should or should not play a central role (or any role) in it. The other is that, similarly, predictive coding does not (or should not) imply anything in particular about the issue of classicism versus 4E (embodied, embedded, enactive, extended) cognition. I will, in turn, explain these two simple points.

First, let us consider the issue of representation within the framework of non-classical cognitive science. It makes sense to structure the multitude of such views in some meaningful way. Following Gallagher (2017) we could order these views according as how committed they are to eliminating representations and how anti-individualistic they are, that is, to what extent the organism is considered as sufficient for cognition and mentality. The extant views range from ones that are very close to classicism in both taking representations as necessary to cognition and taking the organism as the ultimate unit of analysis to ones that pride themselves in being so radical as to eliminate representations completely and hypothesize that the ultimate unit of analysis is the causal loop between organism and its environment or niche. What, then, to make of the idea that a viable alternative to the metaphor of receptor coding could be 4E cognition, given this diversity of 4E views? It looks like the only versions that could serve that purpose are the most radical ones on the above-mentioned two-axis classification. Indeed, Brette explicitly states that one crucial problem with receptor coding is its inadequacy in emulating the organism-environment circular causal loop, at least when it comes to perception; more precisely, given that this loop (reafference) is an ultimate unit of analysis of cognition, it does not make sense to posit receptor coding as a first stage (afference) in perceptual processing.

This idea is coherent, but one wonders, then, why receptors would even get be discussed at all. If the reafference and the continuous circular causal loop of organism-environment interaction is truly the ultimate unit of analysis, then there is nothing special about the receptors to consider, or about any other part of the nervous system for that matter. Then it looks as though the initial problem was not really about *receptor* coding in particular, but about anything like computational processes that are at a lower level than the organism-environment loop. This is not an alternative to receptor *coding*, but an alternative to the receptors themselves – they would cease to have any theoretical role in explaining cognition or perception. Needless to say, one is not forced in any way to take that route even if one is skeptical about the metaphor of coding. On the contrary, one could even conider the PNS (including the receptors) as an *essential* component of the neural realizers of any perceptual or sensory state (Aranyosi 2013) or go even as far as to hypothesize that neuron populations at the receptor level are performing *computations* (Pruszynski & Johansson 2014).

Turning now to predictive coding in particular, contrary to Brette's criticism, it does not necessarily imply intellectualism. It is true that the most interesting and radical way of thinking about predictive coding is one according to which it is rather a "Kantian" *rival* to 4E cognition (e.g. the view defended in Hohwy 2013 and especially in Hohwy 2016), but it is not the only game in town. Indeed, when it comes to perception, there is no reason to think that predictive coding is uncongenial to 4E cognition. On the contrary, as Orlandi (2018, p. 2368) observes, "PCP [*predictive coding approach to perception*] was initially developed in cognitive science in the field of active vision, and it was thought to be good news for proponents of ecological and embodied understandings of perception (Rao and Ballard 1999). It is a curious development that it would be taken up by proponents of more intellectualist accounts." The basic point to emphasize here is that the central tenets of predictive coding (generative models, prediction error minimization, free energy principle) are really about a model of information *communication*; the issue of whether that communication chain is present wholly within the brain, as Hohwy thinks, or spans across the brain-body or even across the body-environment frontier (c.f. Kirchhoff 2015) is orthogonal to the issue of whether the model is adequate for perception or cognition in general. Hence, there is nothing intrinsically uncongenial to more embodied and embedded views of the mind in the idea of predictive coding.

To sum up, while I think Brette points out some significant shortcomings of the metaphor of receptor coding, it is unclear whether an alternative to it, in guise of what he seems to think this would involve, is forthcoming anytime soon.

Beyond Neural Coding? Lessons from Perceptual Control Theory

Xerxes D. Arsiwalla[a], Ruben Moreno Bote[b,c] and Paul Verschure[a,d]

[a]Institute for Bioengineering of Catalonia & Barcelona Institute for Science and Technology, 08019 Barcelona, Spain; [b]Center for Brain and Cognition, Department of Information and Communication Technologies, Universitat Pompeu Fabra, 08018 Barcelona, Spain; [c]Serra Húnter Fellow Programme, Universitat Pompeu Fabra, 08018 Barcelona, Spain and [d]Catalan Institute for Advanced Studies, 08010 Barcelona, Spain.
x.d.arsiwalla@gmail.com https://specs-lab.com ruben.moreno@upf.edu https://www.upf.edu/web/tcn pverschure@ibecbarcelona.eu https://specs-lab.com

doi:10.1017/S0140525X19001432, e217

Abstract

Pointing to similarities between challenges encountered in today's neural coding and twentieth-century behaviorism, we draw attention to lessons learned from resolving the latter. In particular, Perceptual Control Theory posits behavior as a closed-loop control process with immediate and teleological causes. With two examples, we illustrate how these ideas may also address challenges facing current neural coding paradigms.

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

It is noteworthy that many of the challenges to today's neural coding paradigms, pointed out in Brette's target article, are strikingly similar to problems encountered with twentieth-century behaviorism. Other authors have also alluded to this correspondence (Fiorillo et al. 2014; Gomez-Marin 2017). Gomez-Marin comments that once it became possible to look inside neural tissue, the philosophical essence of behaviorism made its way back. Behavior was once again relegated to linear responses, but this time to internal causes. According to Powers (1973b), behavior is control of the animal, by the animal, and should be studied as a circular process from the perspective of the animal, including both immediate and teleological causes. In other words, once one knows the "inside," would one really know everything on the "outside"? The issue seems to be with what is meant by causation in these paradigms. Admittedly, both behaviorism and coding do not consider circular causation, nor do they address teleological aspects of causation. Circular causation also features prominently in the "enactivist" philosophy of mind, where an organism's action and perception are constantly shaped by mutual interaction with its environment (Varela et al. 1991; Verschure et al. 2003). The other problem seems to be the way information theory is used. Fiorillo et al. (2014) make the case for a shift in perspective, where information conditioned on the neuron's biophysics, rather than the experimenter's knowledge, does away with the need for a "neural code" (arguing that encoding/decoding only makes sense from the perspective of an external observer). These works substantiate Brette's main argument, claiming that the implicit definition of "code" used in neural paradigms does not encompass aspects of causation or representation relevant for bridging brain and behavior.

There are, however, lessons we can learn from behaviorism. More specifically, solutions to those problems might also prove insightful today for addressing difficulties encountered in current neuroscience paradigms. In particular, Perceptual Control Theory (PCT), championed by William Powers (1973b), was one such response to behaviorism. PCT originated from early cybernetics, which was concerned with control and autonomy in living organisms. PCT posits that behavior is the process of closed-loop control of what the animal senses, rather than a linear causal response to stimuli. The main insight of PCT was that autonomous goal-directed behavior necessitates a hierarchical control architecture, where higher-level controllers are coupled to lower-level controllers such that the output of one layer provides a reference to the next. An organism performs actions to cancel the effects of disturbances in what it senses, to achieve intended perceptual consequences. Reference signals across this hierarchy constitute immediate or distal goals. The specification of goals for achieving intended consequences constitutes purpose. Ultimate purposes are assumed to be intrinsic, tied to survival drives.

How does this link to neural coding? The important point is that neural coding theories are also trying to explain behavior. However, they attempt to do so by anchoring on linear input–output neural mechanisms, akin to a "switchboard" model of behavior (Powers). The alternative is that neural activity influences actions and actions influence neural activity in terms of what the animal perceives. Behavior is thus a circular control process. This solicits an explanation of how and why the animal itself controls the "switchboard." If neural coding approaches ultimately seek to explain perception and action, then ideas from PCT suggest ways to progress beyond metaphors. Namely, PCT calls for a process-theoretic view of the brain-body-environment system, where circular causation is implemented through hierarchical feedback control.

How does control theory address challenges that neural coding theories face? For this, we now turn to control architectures built upon a hierarchy of forward models. These offer a viable solution to closed-loop adaptive and anticipatory processes. The forward models we refer to are internal models acquired during learning and development. These are akin to the physics and psychology engines discussed in Lake et al. (2017). Let us point to two specific examples, where systems-level control architectures with forward models offer the type of closed-loop causal explanations mentioned above. The first example comes from cerebellar motor control. Herreros et al. (2016) and Maffei et al. (2017) have proposed an anticipatory control scheme involving the vestibular system, where the cerebellum implements a forward model of the motor system being controlled. The model generates anticipatory adjustments to counteract postural and equilibrium disturbances during voluntary movements. It does so by learning to anticipate counterfactual errors in motor action given sensory stimuli and an internal model of the motor system. This closed-loop control architecture has been proposed to model eye-blink conditioning, vestibulo-ocular reflexes, and visual tracking, all involving cerebellar circuits. Physiologically, this implies that timing-dependent plasticity rules of Purkinje cell synapses implement a model of the motor system being controlled by that cerebellar microcircuit (Suvrathan et al. 2016). Our second example refers to the hierarchical mirror system identified in the brain (Gazzola & Keysers 2009). This extends across several brain regions including the motor, somatosensory, and gustatory areas. These systems have been shown to hierarchically implement internal forward and inverse models relating to one's own sensations. The mirror system projects outputs of these self-models upon others during social interactions. This hierarchy of forward and inverse models has been used to explain empathy, somatic sensations in others, and emotions in social cognition (Keysers et al. 2010). Both these examples illustrate the role of control and internal models in brain and behavior.

In closing, one would agree that encoding and decoding of experimental variables in behavioral paradigms are valuable epistemological constructs for the experimenter. If "goals" and "purposes" are necessary to describe behavior emerging from dynamical systems engaged in hierarchical control (Powers), then neuronal coding protocols might be useful tools to identify precisely those variables that define the underlying closed-loop dynamical system. However, it is also true that one ought to refrain from the fallacy of extending conditional epistemic descriptors to ontological explanations of brain and behavior.

Generative models as parsimonious descriptions of sensorimotor loops

Manuel Baltieri and Christopher L. Buckley

EASY Group – Sussex Neuroscience, Department of Informatics, University of Sussex, Brighton BN1 9RH, United Kingdom.
m.baltieri@sussex.ac.uk https://manuelbaltieri.com
c.l.buckley@sussex.ac.uk https://christopherlbuckley.com/

doi:10.1017/S0140525X19001353, e218

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

Abstract

The Bayesian brain hypothesis, predictive processing, and variational free energy minimisation are typically used to describe perceptual processes based on accurate generative models of the world. However, generative models need not be veridical representations of the environment. We suggest that they can (and should) be used to describe sensorimotor relationships relevant for behaviour rather than precise accounts of the world.

In the target article, Brette questions the use of the neural coding metaphor in the neurosciences. One of the main arguments is related to the criticism of approaches that overemphasise the role of perception as opposed to motor control for accounts of cognition. As suggested by Brette, while the sense-model-plan-act paradigm has long been criticised (Brooks 1991b), it still survives in modern approaches to neuroscience. He then extends this criticism to the notion of efficient coding and its most recent heir, predictive coding. Specifically, he argues that models that describe perception as a process of minimising redundancy (efficient coding) or prediction error (predictive coding) between incoming sensations and a generative model of sensory data are often used to ascribe the "goal" of minimising redundancy/error to an agent, to account for the rich repertoire of behaviours of living organisms. The shortcomings of this approach have previously been discussed, and resulted in the emergence of problems such as the dark-room paradox (Friston et al. 2012): why should agents show complex behaviour when, to minimise redundancy/prediction error, they could easily find the most predictable state in the world, for example, a dark room?

Recently, the ideas of predictive coding/processing have been extended to include accounts of action, for example, *active inference*. On this account, while perception is a process of changing predictions to better account for sensory data, action is a process of changing the world to better meet predictions. Normative behaviour then arises in this framework if an agent predicts rewarding states (Friston et al. 2012). Active inference moves the goal of cognitive agents from inferring properties of sensory data to acting in order to meet their goals. This extension of predictive coding is thus not in conflict with the ideas in the target article but, rather, directly supports them.

Furthermore, adaptive behaviour can emerge even if generative models are far divorced from a veridical representation of the environment. These *action-oriented* generative models, described in the context of *radical* predictive processing (Clark 2015), can operate on the basis of linear approximations of world dynamics (Baltieri & Buckley 2019b) or simple sensorimotor couplings rather than objective properties of the world (Baltieri & Buckley 2017). These ideas are derived from 4E (embodied, enactive, embedded, and extended) theories of cognition which have long sought to address different issues of computationalism (Newen et al. 2018), including the misuse of properties metaphors such as neural coding.

Thus, the idea of predictive processing/coding that the author describes, based on models that generate accurate representations of observed data and inherited from statistics and machine learning, is not the only game in town. Instead, we advocate for ideas of predictive coding that include generative models that are parsimonious descriptions of sensorimotor contingencies (Baltieri & Buckley 2017). This may sound like an unnecessary stretch of the definition of a generative model (Bruineberg et al. 2018) but, we argue, is far from being just a semantic argument. The mathematical definition of these models is still entirely consistent with the more familiar notion of "generative" models and constitutes a valuable framework for the modelling of sensorimotor loops because of its strong and established relationships with (optimal) control theory (Todorov 2009). In this context, "inference" can be best understood as a process of estimating actions necessary to attain future goals, *generating* expectations of desired states of affairs (rather than objective truths about the world) that are brought into existence by means of active behaviour.

In active inference, generative models further diverge in some fundamental ways from the more traditional ideal-observer-based forward models used in the context of motor neuroscience (Friston 2011) and discussed in the target article. Forward models rely largely on a Kalman-like approach where all the variables affecting a system (parameters and inputs or causes, in a state-space formulation sense) and observations of said system (outputs) are available for an ideal observer to infer is latent states. On the other hand, generative models in active inference explicitly take the perspective of an agent into account. This includes discarding the idea that all causes affecting observations are known to an agent, suggesting instead the presence of approximate mechanisms to implement actions purely based on incoming sensations (not their estimates or "predictions") (Baltieri & Buckley 2019a).

The target article further argues that theories based on Shannon communication theory (including efficient and predictive coding, but more in general all notions of neural coding) cannot in principle explain meaning in biological systems, as this definition explicitly precludes a study of the semantics of information. We agree with Brette that notions of semantic information (with meaning for an agent, not necessarily for an experimenter) are largely neglected in neuroscience but his claim seems to overlook some of the efforts made to extend Shannon's work (Kolchinsky & Wolpert 2018). To understand the implications of the value of information in biological systems, we suggest not to ignore frameworks based on Shannon information, but rather to further look into their connections to causal frameworks. In this light, extensions of predictive coding models such as active inference based on Bayesian networks could, in principle, describe biological agents as intervening in the world using motor actions to learn causal relationships between their actions and their sensations (Hohwy 2013), that is, the "subjective physics of the world" described by Brette.

The neural coding metaphor has long exhausted its appeal to explain cognition. This is because metaphors that attempt to describe living organisms as passively gathering and representing information from the environment cannot account for the complexity of the behaviour of biological systems. Similarly, we believe that a narrow interpretation of generative models as veridical world models is also a misdirection for cognitive science. Instead, we argue that the generative models at the heart of active inference, which are not to be seen as accurate maps of the world but as descriptions of useful information and desires for an agent in the form of parsimonious actions/percepts relationships, will be a valuable tool for the study of cognitive systems.

Acknowledgments. This work was supported in part by BBSRC Grant BB/P022197/1. MB thanks Martin Biehl and Simon McGregor for insightful discussions.

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

Codes, functions, and causes: A critique of Brette's conceptual analysis of coding

David Barack and Andrew Jaegle

Jerome L. Greene Science Center, Columbia University, New York, NY 10027; DeepMind, London N1C 4AG, United Kingdom.
dbarack@gmail.com drewjaegle@google.com www.deepmind.com

doi:10.1017/S0140525X19001407, e219

Abstract

Brette argues that coding as a concept is inappropriate for explanations of neurocognitive phenomena. Here, we argue that Brette's conceptual analysis mischaracterizes the structure of causal claims in coding and other forms of analysis-by-decomposition. We argue that analyses of this form are permissible and conceptually coherent and offer essential tools for building and developing models of neurocognitive systems like the brain.

Brette argues that coding is an inappropriate concept for explanations of neurocognitive phenomena. Brette identifies three properties of coding: correspondence, representation, and causality. Brette grants correspondence but rejects both representation and causality for the neural code. Although we disagree with his analyses of representation and causality, we limit our critique to the latter.

Brette's argument against causality focuses on two points. First, coding assumes that the parts of a cognitive system have separate functions. However, Brette claims that function cannot be attributed to the brain's parts. Second, coding implies linear causality for the brain. Brette argues instead that the brain features circular, coupled causality.

We argue that functions can be attributed to the parts of brains and, though brains are dynamical systems with circular causality, linear causality may still apply. We contend that the rejection of functions for parts of the brain constitutes a direct attack on the nature of explanation in cognitive neuroscience. Furthermore, the causality claim commits a category mistake, as the linear structure of the concept need not be mimicked by the causal structure of the brain. Finally, linear approximations are immensely successful in neuroscientific explanations.

Brette first argues against the assignment of decoding and encoding functions to parts of the brain. Such assignment requires the analysis of behavior "independently of the system in which the neurons are embedded" (sect. 4.1, para. 2). But such an analysis "determines a neural code that … depends neither on the goals of the animals nor on the effect of spikes on … actions" (sect. 4.1, para. 6). He concludes that the analysis of the brain by decomposition cannot proceed because "function can be meaningfully ascribed to the organism as a system, but not … to the components of this system" (sect. 4.1, para. 6). As such coding functions are defined independently of the organism's goals, Brette rejects the possibility of assigning coding functions to parts of the brain.

Brette's analysis relies on several misleading claims. First, some notions of function that do not rely on goals, such as causal role functions (Cummins 1975; Walsh & Ariew 1996), can be attributed to parts of organisms. Second, nothing about goals prevents function ascription to the organism's parts while permitting function ascription to the whole organism. As part of a larger system, the function of the part could share the goal of the organism. Indeed, this is typical for biology, where functions are often assigned to organs – such as the circulation of blood for the heart or cleaning the blood of toxins for the kidneys – even though the goals of these functions might be for the organism. Third, encoding and decoding can take into account variables relevant to the organism's biological fitness (Rice 2015), and goal-oriented functions are defined and ascribed with respect to those fitness-relevant variables.

Furthermore, we contend that coding and other analysis-by-decomposition models are indispensable to explanations of brain function that integrate with psychology. These models break down psychological phenomena into subfunctions for explanation (cf. Dennett 1981; Lycan 1981; Marr 1982a). Decomposition requires that the psychological properties of subfunctions be reduced or removed when ascribing those subfunctions to component parts and proceeds recursively with finer decompositions with fewer psychological properties. At the lowest levels, functional descriptions are completely bare of psychology, yielding a reduction to neuroscience. Coding is a perfect example of such decomposition. The concept of coding implies encoding and decoding functions with reduced intentional implications, as those functions are grounded purely in probabilistic terms (Shannon & Weaver 1963). While message contents still need to be determined, coding analyses of systems like the brain can result in parts that carry weaker intentional properties.

Brette next argues that the causality implied by coding does not apply to the brain. The neural code has linear causality (viz., input → encoder → decode → output), whereas the brain possesses circular, coupled causality. The brain is like a tent, where "different elements are co-determined…. In addition to the coupling of neurons, the brain itself is coupled to its environment; that is, there is circular and not linear causality" (sect. 4.4, para. 2). Linear causality refers to temporally sequential, causally related pairs of states, whereas tentlike causality refers to simultaneous, jointly causally related sets of states. In short, the causal structure implied by the neural code fails to match the causal structure of the brain.

We first note that tentlike causality is consistent with linear causality. Linear encoding-decoding relationships between each pair of elements are consistent with an overall picture of a circular, coupled causal system. Indeed, Brette arguably commits a category mistake (Ryle 1949): while the conceptual analysis of coding involves a linear structure (Shannon & Weaver 1963), the implementations of encoding and decoding functions need not. This category mistake is further illustrated by Brette's claim that coding structures imply the "discrete temporality of an algorithm, … disconnected from physical time…. But … dynamical systems cannot in general be mapped to algorithmic descriptions" (sect. 4.4, para. 1). However, nothing about the coding metaphor entails discrete time, and one variable may encode another in the sense of carrying Shannon information without being part of an algorithm.

Brette's argument also ignores the utility of linearity and discrete time for analyzing complex systems. Continuous-time systems can be well approximated by discrete-time systems (Oppenheim & Schafer 2013), and many equivalences exist between dynamical systems with circular causality and approximators with iterated linear causality (e.g., Funahashi et al. 1993; Schäfer & Zimmerman 2007). Furthermore, systems with tent-like causality can be approximated by linear causal elements, such as by iteratively computing local relationships (e.g., Geman & Geman 1984; Roth & Black 2005; Van Den Oord et al. 2016). Deep learning research uses discrete-

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

time architectures to model many forms of continuous behavior as well (e.g., handwriting [Graves 2013], speech synthesis [Van Den Oord et al. 2016], video prediction [Srivastava et al. 2015], robotic control [Levine et al. 2016], humanoid running [Heess et al. 2017]). Hence, complicated, ethologically relevant behavior can be analyzed with linear causality and discrete time, and this analysis is likely to remain a crucial part of building and designing models of this behavior (Santoro et al. 2019).

From the "coding metaphor" to a theory of representation

Jonathan Birch[a] and Joulia Smortchkova[b]

[a]Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, London WC2A 2AE, United Kingdom and [b]Faculty of Philosophy, Radcliffe Humanities, Radcliffe Observatory Quarter, University of Oxford, Oxford OX2 6GG, United Kingdom.
j.birch2@lse.ac.uk joulia.smortchkova@philosophy.ox.ac.uk
http://personal.lse.ac.uk/birchj1 https://jouliasmortchkova.wordpress.com

doi:10.1017/S0140525X19001456, e220

Abstract

Brette highlights a conceptual problem in contemporary neuroscience: Loose talk of "coding" sometimes leads to a conflation of the distinction between representing and merely detecting a property. The solution is to replace casual talk of "coding" with an explicit, demanding set of conditions for neural representation. Various theories of this general type can be found in the philosophical literature.

Although Brette's official target is the "coding metaphor," he often seems more broadly sceptical of the idea that patterns of neural activity can represent properties of the external environment. His scepticism is motivated by examples where, although a neuron's activity is sensitive to an environmental property and carries correlational information about that property (i.e., information in Shannon's sense), this close dependence is a fragile artefact of a particular experimental setup, prone to disappear as soon as the subject leaves the lab and re-enters the wider world.

He gives us several cautionary tales along these lines. For example, a single cone cell carries correlational information about wavelength in a laboratory setup that fixes the intensity of the light, but the correlation breaks down when the intensity is allowed to vary. A neuron in the medial superior olive may carry correlational information about the location of a sound in a laboratory setup that holds fixed intensity and frequency, but intensity and frequency will not be fixed outside the lab.

Brette argues that the use of the term *coding* to describe these fragile correlations is misleading: it obscures their dependence on a specific laboratory context, suggesting instead that the correlation is robust and general. He also laments a particular kind of slip: the existence of a fragile correlation leads to talk of coding, and this leads to an assumption that neural activity represents the detected property, where "representing" goes beyond merely "correlating with" and implies semantic content.

The article brings out a real conceptual problem in contemporary neuroscience: representing a feature in the environment is sometimes considered synonymous with detecting that feature in a controlled setting. Loose talk of "coding" may be part of this problem, if it encourages researchers to conflate the distinction between detection and representation. But is the only alternative abandoning all talk of representation in the brain?

A different approach is to set out an explicit and demanding set of conditions under which a pattern of neural activity genuinely represents an environmental property. This approach has been pursued in some depth in the philosophical literature on mental representation. Although philosophical accounts of mental representation are many and various, none simply identifies representation with correlation. Typical additional requirements include the following: (a) the correlation between the detected feature and the pattern of neural activity is exploited in further processing; (b) this further processing ends up guiding action and explaining success; and (c) misrepresentation is possible (e.g., when neurons fire in the absence of the stimulus they evolved to detect), and this helps explain unsuccessful action. These criteria have been discussed under different guises in the philosophical literature on causal and teleosemantic theories of mental content (Dretske 1981; Millikan 1984; Neander 2017; Shea 2018). With a theory of this general type in hand (such as Shea's or Neander's), there is room for a serious empirical debate about whether a given pattern of neural activity represents the property it detects.

As an illustrative example, consider mirror neurons: neurons that fire both during motor action production and during observation of motor acts in others. These neurons were first discovered in area F5 and in the inferior parietal lobule of the monkey and later found in functionally equivalent areas of the human brain (Fabbri-Destro & Rizzolatti 2008). There is no doubt that their activity correlates under controlled conditions with properties of a variety of motor acts, such as grasping movements. They belong to an information-processing route that involves other brain regions and they receive information from higher visual areas. There are also well-documented contextual effects: for example, some (but not all) neurons that fire in response to a perceived grasping movement will also fire if the movement is partially occluded by a screen, but only if the agent previously saw a graspable object hidden behind the screen (Umiltà et al. 2001).

Here we have systematic correlation, inviting casual talk of neurons "coding" for motor acts. Yet it would be hasty to infer that the neurons represent the properties they detect. There is a long-running debate about their function in cognition and about what they represent (Cook et al. 2014). To settle that, we need to address a package of further questions: Do mirror neurons detect the same type of property in a variety of experimental settings and with different techniques? Are the correlations with properties of motor acts exploited in further processing? Do they guide successful action? Are there cases in which unsuccessful action can be explained by the mirror neurons misrepresenting a property of a motor act? As long as clear answers to these questions remain elusive, we should remain agnostic on the question of what features of the external world mirror neurons represent. However, the fact these questions are difficult to answer is no reason to doubt the validity and value of the concept of representation for understanding the brain. On the contrary, the concept of representation is exactly what we need to formulate and test hypotheses about the function of mirror neurons. To call for the elimination of the concept would be akin to arguing that, because it is often hard to tell

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

whether or not a trait is an adaptation, we should eliminate the concept of "adaptation" from evolutionary biology.

In sum, Brette correctly highlights the shortcomings of the "coding metaphor" as a theory of representation in the brain. His article points to the need to replace fairly loose talk of "coding" with more precisely articulated concepts of feature detection, correlation, and representation. The concept of representation is especially onerous and should be used with great care. But to eliminate the concept altogether would be throwing out the baby with the bathwater.

Modest and immodest neural codes: Can there be modest codes?

Rosa Cao[a] and Charles Rathkopf[b]

[a]Department of Philosophy, Stanford University, Stanford, CA 94305 and [b]Institute for Neuroscience and Medicine, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany.
rosacao@stanford.edu c.rathkopf@fz-juelich.de
https://philosophy.stanford.edu/people/rosa-cao http://charlesrathkopf.net/

doi:10.1017/S0140525X19001420, e221

Abstract

We argue that Brette's arguments, or some variation on them, work only against the immodest codes imputed by neuroscientists to the signals they study; they do not tell against "modest" codes, which may be learned by neurons themselves. Still, caution is warranted: modest neural codes likely lead to only modest explanatory gains.

Coding in the context of human communication involves using one set of symbols to stand in for another. A codebook specifies the appropriate interpretation of symbols in the code by mapping them to *already meaningful* representational entities (e.g., words, letters, pictures). How can this notion be extended to signals exchanged between simpler senders and receivers that have no pre-existing linguistic or representational practice?

Brette thinks this cannot be done for neural signals for two reasons. First, what content *is* assigned in practice depends as much on the experimenter's interpretations and interests as it does on the system being studied. And second, signalers in the brain couldn't *possibly* have access to the external facts needed to establish meanings for the symbols that they use.

On the first point, he is absolutely right. Brette's concerns about the context-boundedness of neural codes are close kin to well-known philosophical worries about the meaning of biological signals: Mere correlations are too permissive to attach unique contents to signals; what is needed in addition is something like a normative *function* for the signal – a target relative to which its performance may be judged, either by evolution, punishment, or some other mechanism (Dretske 1994). But it's commonly agreed that learning- or evolution-based normative functions can never be so precise as to allow us to use them to specify perfectly determinate, non-disjunctive contents. Some have thought this fatal for the project of naturalizing representations, but others argue that whatever indeterminacy we end up with is a feature, not a bug: Biological function is somewhat indeterminate, and so too is meaning; there is no further fact of the matter to worry us (Fodor 1990; Neander 1995; Papineau 2003).

Neuroscientists, meanwhile, sidestep these problems by directly (if often implicitly) stipulating the relevant functions and representational targets themselves. For example, as oriented bars are used in the experiment, oriented bars must be what the V1 neurons represent. But this is just to build into their experimental design and interpretation what they think the relevant neural function must be and, thereby, to end up partially stipulating the meaning they claim to have discovered.

It is Brette's second point that we want to focus on. *If* we insist that neural signals have the same rich semantic properties as conventional human symbols, then indeed there seems to be a puzzle as to how neurons could learn such meanings, given the stark differences between the world as scientists see it and what neurons themselves have access to. But why *must* we insist on such a demanding notion of code? To require an *independent* reservoir of meaning, from which we can draw semantic labels to stick on neural signals is not just a neuroscientist's fantasy, but a fantasy *tout court*. Human language itself has no such reservoir to appeal to – ultimately, our symbols acquire meaning by virtue of our conventions and practices with respect to them. Insofar as action policies can be established among parts of the brain that need to coordinate their activities with each other and the outside world, why shouldn't some neural signals likewise acquire meanings by virtue of *their* action policies?

Brette thinks it is impossible for neural signals to be interpreted by the brain in the necessary ways. But it seems to us that the key ingredient is available, at least for a modest notion of encoding. What is required to develop an effective codebook is just the capacity to *learn* from ongoing interaction with the world, and as Brette points out (ironically, in defense of the opposite position), plasticity is one of the brain's most prominent and unavoidable characteristics. This is bad if you think that codes must be Platonic and unchanging, but good if you agree with us that codes can be learned – and moreover that brains have evolved to do just that.

Philosophers have developed mathematical models showing how action policies can endow bare causal commerce with meaning (Skyrms 2010). As a consequence, and *contra* Brette, signals might well acquire something plausibly meaning-like as a consequence of the functional role they gradually learn to play in the overall economy of the brain.

For Brette's blind iguana to develop a code, its neurons need only learn an action policy exploiting the correlation between location and sound, built from dynamic feedback (again, a feature that Brette emphasizes) with the environment. Internal signals will then be message-like in the sense that they help neurons coordinate their activities with each other to produce coherent responses to particular environmental stimuli (Rathkopf 2017).

Failures of the policy can then count as failures of representation, and not merely a breakdown in the causal or correlational structure of the system. Why? *Because* the neurons ended up with *these* action policies (and not some other ones) *by virtue of* the success of the responses thus produced, successes which account for the policies coming to be stabilized in the first place (Cao 2012; Millikan 1984; Shea 2018). That is the reasoning that leads us to say that the *function* of these neurons is to produce the effects of these action policies, in response to these environmental cues.

Of course human practices with respect to the symbols we exchange with each other are more flexible and more sophisticated/

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

articulated than those among neurons. That flexibility and sophistication eliminates some indeterminacy in what we mean, but not all. Neural signals are likely much less determinate, which explains, in turn, why neuroscientists are both able and tempted to affix their own interpretations, informed by *their* explanatory interests.

We sympathize with Brette's suspicion of strong neural codes when they require illicitly projecting our human conceptual scheme into the brain's inner workings. This doesn't mean that we should never attribute contents when convenient, just that we should be explicit when doing so, avoiding "semantic drift." A more modest notion of coding may help us understand the directedness of the brain's coordinating activities, while avoiding anthropocentric contents, because the meanings that arise from such a learned, action-based code are not ours, but the brain's. It would be a mistake to identify those neural contents with the psychologically salient meanings that we ourselves experience. Thus, while modest codes may be available, their explanatory payoffs are likely to be correspondingly modest, and so either way, caution about the coding metaphor is warranted.

A sensorimotor alternative to coding is possible

Paul Cisek

Department of Neuroscience, University of Montréal, Montréal, QC H3C 3J7, Canada.
paul.cisek@umontreal.ca www.cisek.org/pavel

doi:10.1017/S0140525X19001468, e222

Abstract

If we abandon the coding metaphor in favor of models of the full behavioral loop, we need a way to dissect that loop into understandable pieces. I suggest that evolutionary data provide a solution. We can subdivide behavior into parallel sensorimotor subsystems by following the phylogenetic history of how those systems differentiated and specialized during our evolution, leading to promising ways of re-interpreting neural activity within the context of its pragmatic role in mediating interaction.

Brette provides a timely critique of the coding metaphor and suggests the alternative lies in models of the full sensorimotor loop (Ashby 1965; Brooks 1991a; Cisek 1999; Clark 1997; Dewey 1896; Gibson 1979; Hendriks-Jansen 1996; Piaget 1963; Powers 1973a). Indeed, the fundamental task for an organism is not to encode the environment but to complement its dynamics so the entire brain-body-environment system flows toward states supporting the organism's survival and away from those that don't.

However, the full sensorimotor loop is so complex that understanding it all is a daunting task. This is partly why the coding metaphor is so pervasive – it offers a tempting method to delineate subsystems within the loop, each with defined inputs and outputs, which can then be studied experimentally. But if splitting the loop into sensory, motor, and cognitive processes leads to artificial borders and flawed notions of coding, then how else can we subdivide the large question of behavior into smaller and more manageable questions?

One possible answer lies in evolution. The brain evolved through a long series of modifications that gradually elaborated older systems into newer ones, and the history of these modifications can be reconstructed from comparative data. With that history as a guide, one can develop theories of behavior through a stepwise process that provides a different way of dissecting behavior (Cisek 2019; Grossberg 1978). Instead of breaking the sensorimotor loop into modules such as "perception" and "decision-making," we can consider how distinctions between control mechanisms emerged in evolution through differentiation and specialization of ancestral systems.

Cisek (2019) briefly summarizes our evolutionary history as the continuous expansion of the behavioral repertoire by differentiating control mechanisms and extending them into progressively more abstract domains (Hendriks-Jansen 1996; Piaget 1963). These mechanisms operate as sensorimotor loops that ensure the brain-body-environment system flows toward desirable states. As spikes are a means of directing that flow, their activity perforce corresponds to aspects of the world, but also to the organism's needs and its policies for meeting those needs. We could call these a "pragmatic representation" – activity that doesn't describe the world but instead mediates interaction with it. While it may seem inappropriate to call this a "representation," the term allows one to consider pragmatic representations as one end of a continuum at the other end of which lie "descriptive representations" – which still serve a pragmatic role but have at least partially detached from internal context. We can then use these concepts to interpret neural organization.

The mammalian neocortex consists of two sheets: a dorsomedial sheet sensitive to topographic space, and a ventrolateral sheet that is nontopographic (Finlay & Uchiyama 2015). Dorsomedial neocortex is organized as a set of circuits for controlling different aspects of the animal's behavioral repertoire (Graziano 2016; Kaas & Stepniewska 2016). Each of these contains a map of potential actions that compete for execution, biased by signals from other regions such as the basal ganglia and ventrolateral cortical regions, including temporal and orbitofrontal areas (Cisek & Thura 2018; Yoo & Hayden 2018). In this view, behavior consists of a constant competition between potential actions present in the environment (Cisek 2007; Cisek & Kalaska 2010).

For example, let's consider the primate dorsal premotor cortex (PMd), long associated with planning and control of reaching (Wise 1985). PMd neurons are selective to reach direction, speed, and amplitude, but their tuning appears to change over time and with intended speed (Churchland & Shenoy 2007), and they are modulated by attention (Lebedev & Wise 2001), speed-accuracy trade-offs (Thura & Cisek 2016), expected probability (Thura & Cisek 2014), and magnitude of rewards (Pastor-Bernier & Cisek 2011) in a manner that depends on the angular distance between reach targets. This makes decoding PMd impossible unless one knows every detail of the experimental context, as Brette describes for sensory systems. However, if we interpret PMd as part of a dynamical system that guides the arm to a desirable target (Churchland et al. 2010), then these findings make good sense. Before movement, a target must be selected, which can be accomplished through competition among recurrently connected neurons (Cisek 2006; Erlhagen & Schoner 2002; Grossberg 1973) biased by any factors that bear upon the animal's choice. In this view, PMd activity is amplified when monkeys make hasty guesses (Thura & Cisek 2016) not because it "encodes hastiness," but because increased tonic arousal accelerates winner-take-all selection in recurrent networks (Grossberg 1973). Once movement starts, these correlations vanish, and activity becomes specific to a given movement. Thus, PMd behaves just as one would expect from a

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

dynamical system for specifying, selecting, and guiding actions, even if those actions cannot be decoded from its activity.

While dorsomedial neocortex may be specifying competing potential actions, ventrolateral regions may be detecting cues to help bias that competition – what ethologists call "key stimuli" (Hinde 1966). Pragmatically most important is that the activity helps to select the right thing to do in the world, not to describe the world. For example, activity related to visual information coming from an apple should be amplified by hunger and reduced by cues that indicate a predator.

In a few special cases, however, a pragmatic representation may gradually detach from its context dependence and become increasingly "descriptive." If interaction is based on exploiting affordances in the world (Gibson 1979), then it is important that those affordances are often attached to specific objects and places. Thus, there is a pragmatic role for categorical processes in ventrolateral cortex that group key stimuli into activity that classifies objects, especially those of most relevance for a given species' behavior (Leopold et al. 2017). And as control is further extended to complex social interactions, so it becomes even more useful to construct encapsulated "symbols" that can mediate coupling in the agent-agent system. The words on this page are an example.

In summary, I agree with Brette's critique of the coding metaphor and his suggestion that we replace it with closed-loop sensorimotor control. However, his message could be neglected simply because abandoning coding may seem like abandoning the possibility of subdividing the large problem of behavior into manageable subproblems. Fortunately, evolution provides us with an alternative approach. Instead of subdividing behavior into information processing modules resulting from the history of psychological theories, we can resynthesize those theories by following a different kind of history: that of our own evolution.

Acknowledgments. The author is supported by the Canadian Institutes of Health Research (MOP-102662), the Natural Sciences and Engineering Research Council of Canada (RGPIN/05245), and the Fonds de la recherche en santé du Québec.

Is information theory, or the assumptions that surround it, holding back neuroscience?

Lee de-Wit[a], Vebjørn Ekroll[b], Dietrich Samuel Schwarzkopf[c,d] and Johan Wagemans[e]

[a]Department of Psychology, University of Cambridge, Cambridge CB2 3EB United Kingdom; [b]Department of Psychosocial Science, University of Bergen, 5020 Bergen, Norway; [c]School of Optometry & Vision Science, University of Auckland, Grafton, Auckland 1023, New Zealand; [d]Experimental Psychology, University College London, London WC1N 1PJ, United Kingdom and [e]Laboratory of Experimental Psychology, University of Leuven (K.U. Leuven), B-3000 Leuven, Belgium

lhd26@cam.ac.uk Vebjorn.Ekroll@uib.no s.schwarzkopf@auckland.ac.nz johan.wagemans@kuleuven.be

https://www.psychol.cam.ac.uk/people/lee-de-wit

https://www.uib.no/en/persons/Vebjørn.Ekroll

https://unidirectory.auckland.ac.nz/people/d-schwarzkopf

http://www.gestaltrevision.be/en/about-us/principal-investigator

doi:10.1017/S0140525X19001250, e223

Abstract

The challenges raised in this article are not with information theory per se, but the assumptions surrounding it. Neuroscience isn't sufficiently critical about the appropriate 'receiver' or 'channel', focuses on decoding 'parts', and often relies on a flawed 'veridicality' assumption. If these problematic assumptions were questioned, information theory could be better directed to help us understand how the brain works.

We agree that the target article is right in highlighting some of the sloppiness with which information theory is used in neuroscience, and that in some cases the term *code* is used to describe the relationship between neural activity and sensory input, when actually the term *correlate* would be more appropriate. We have previously argued that neuroscience all too often focuses on what an experimenter can decode from neural activity, rather than testing what (or whether) the brain might be able to decode from that activity (de-Wit et al. 2016). Rather than reflecting a fundamental problem with information theory, however (which always clearly acknowledged the importance of a receiver), we argue this actually reflects the way in which "coding" is used too loosely as a metaphor.

Information theory always required a careful consideration of the sender, channel, and receiver. We have previously argued that neuroscience is often too hasty in assuming that the "channel" used by the brain to convey information is the firing rate of neurons when, of course, there are many aspects of neural activity (synchrony, precise timing) which could also be used as the channel over which information is conveyed. Again, it is not the fault of information theory that there is insufficient consideration in neuroimaging regarding the channel by which information might be conveyed. This is particularly significant for strong claims regarding an "absence" of information when using fMRI, which is completely unable to detect information that might be conveyed by precise temporal coding.

One of the other weaknesses highlighted in the target article, namely, a lack of consideration of context effects, also does not reflect a fundamental problem with information theory, but rather a limited theoretical framework within visual neuroscience. We would argue that the Gestalt tradition offers a much clearer insight into some of the challenges faced in processing sensory input, but that modern neuroscience largely tries to side-step these challenges and simply focuses on correlations between "parts" of the input and neural activity. From a Gestalt perspective trying to understand how "parts" might be represented in the brain without also considering how those parts will form together into "wholes" was obviously going to be a limited enterprise from the start. Indeed, even seemingly simple "parts" like edges actually need to considered in context to understand what information neural signals might be conveying (Kogo & Wagemans 2013).

Finally, however, the target article does make a stronger more fundamental argument, that the coding metaphor is wrong because information theory can account only for the reference between information states and objective properties of the world. Here we also agree that perception is not simply a process of "re-presenting" the "objective" properties of the external world. This is a bigger challenge for information theory, but we would argue that this challenge always should have been at the heart of how information theory was used in neuroscience.

Indeed, the focus on understanding the "goals" of different information processing systems was clearly articulated as an

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

important part of Marr's (1982b) levels of analysis, but is often neglected in modern neuroscience. Indeed, much neuroscience starts from an assumption that there is only one version of reality that can be derived from sensory input, and that the job of the visual system is simply to "optimally decode" that representation of reality. A biologically plausible theory of information processing also has to address the goals of the organism (as Marr made clear) to then think about what kinds of representations might be useful to achieve those goals. A frog's visual system may extract transient motion signals to provide information about food/flies; a stickleback fish may represent curved contrast boundaries with specific wavelengths of light to provide information about competing mates; a human visual system will combine different views of the same object to provide information that this is the same object over time. Thus, what different organisms do with sensory evidence will differ depending on their *umwelt* (von Uexküll 1992) and their evolutionary needs, but we would argue it is still useful to think about those perceptual systems as representing the information that is useful for (and subjective to) their goals.

The idea that different representations might be derived from the same message is not an entirely alien concept for information theory. The content of an encrypted message, for example, is dependent on the receiver having the right decryption key. Thus, decoding is always subjective not only to the content from the sender, but to the computations performed by the receiver. Different visual systems will most certainly decode different representations from the same input but, clearly framed in this way, we would argue that information theory is not necessarily the problem here; rather, the problem is many of the assumptions that surround its use.

Whether or not information theory will ultimately prove useful in explaining the activity of the brain is of course an empirical question, but we would argue that more progress could be made with information theory if the problem of perception were conceived more correctly. The target article is right in highlighting many of these problematic assumptions, and we agree we cannot look for neural codes in any sensible way without questioning these assumptions. But we would question whether the problem really lies with information theory, or the lack of a psychologically and biologically plausible account of the information processing challenges our brain has to solve. To paraphrase Mausfeld (2003), if we don't put any substantive psychological theory into the use of information theory in neuroscience, then we can't expect any plausible psychological answers from it.

Abandoning the code metaphor is compatible with semiotic process

Terrence W. Deacon[a]
and Joanna Rączaszek-Leonardi[b]

[a]Department of Anthropology, University of California at Berkeley, Berkeley, CA 94720 and [b]Faculty of Psychology, University of Warsaw, 00-183 Warsaw, Poland.
deacon@berkeley.edu raczasze@psych.uw.edu.pl
http://hill.psych.uw.edu.pl

doi:10.1017/S0140525X19001419, e224

Abstract

We agree with Brette's assessment that the coding metaphor has become more problematic than helpful for theories of brain and cognitive functioning. In an effort to aid in constructing an alternative, we argue that joining the insights from the dynamical systems approach with the semiotic framework of C. S. Peirce can provide a fruitful perspective.

Although some commentators may argue that the code metaphor has been set up as a strawman by Brette, it takes little effort to catalogue its ubiquitous use in the neurosciences over the past half-century. The influence of this conceptual framing has been reinforced by the highly successful technique of recording the spike trains of individual neurons in response to stimulus presentations. Thus, a correlation between specific stimulus features and rapid spike production by a specific neuron is presumed to license the claim that this neuronal activity in some way *encodes* that stimulus feature. Yet correlated neural activity of any of the potentially large number of neurons located anywhere along the path linking the initial registration of the stimulus to a specific neuron thereby caused to become highly active could likewise be understood as encoding that same stimulus.

It must indeed be the case that signal transduction from neuron to neuron is in some way necessary for brain processes to be *about* anything. But the problem with treating correlations or covariant dynamics as the sufficient basis for explaining cognitive or even perceptual functions is that the designated aboutness is only in the eye of the experimenter, not an intrinsic property of neural processes. But is this even a useful heuristic fiction? We agree with Brette's assessment that framing the problem in these terms has become more problematic than helpful, precisely because it is often accepted as an explanation when instead it is merely descriptive.

But having exposed the dangers of employing the code metaphor in cognitive neuroscience, we need a more appropriate alternative. One standard response is to abandon the concept of representation altogether and use only the language of covarying coupled dynamics. But this merely replaces an atomistic mapping relation with a dynamical mapping relation. In this commentary we argue that joining the insights provided by a dynamical systems perspective with the semiotic framework of C. S. Peirce (Hartshorne & Weiss 1931–1963) can provide a middle path between the atomistic reductionism of an encoding paradigm and the "dynamics only" approaches. This is because each of these frameworks addresses key weaknesses in the other. A dynamical systems perspective can help to ground the notion of interpretation that is framed only in formal terms within semiotic theories, whereas a semiotic perspective can help to disentangle the distinctive roles of dynamics and form in dynamical systems theories.

A key concept necessary to bridge these frameworks is the concept of constraint. This is a fundamental concept both in dynamical systems theories and in information theory, which provides a precise analytical tool for characterizing relations of form. Two seminal thinkers in theoretical biology – Howard Pattee (1973) and Michael Polanyi (1968) – have argued that focusing on the relationship between constraints and dynamics is essential if we are to understand the fundamental logic of living processes. They each stress that these complementary aspects of a living process must at the same time be functionally interdependent but physically independent.

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

Thus, Polanyi (1968) argues that the constraints that organize the dynamical processes of life are "Irreducible higher principles [that] are additional to the laws of physics and chemistry" (p. 160). And similarly, Pattee (1997) points out that "Physical laws and semiotic controls require disjoint, complementary modes of conceptualization and description. Laws are global and inexorable. Controls are local and conditional. Life originated with semiotic controls" (p. 9).

For Pattee, "semiotic controls" exist in the form of "non-integrable constraints" that involve a "necessary epistemic cut between the coherent physical dynamics and its rate-independent semiotic description." Importantly, constraints are off-loadable onto artifacts and their structure/form. This facilitates and is necessary for the preservation of constraints across potential changes of dynamics.

Although the Peircean semiotic framework is often treated as though it is a structuralist typology of sign types, it is also compatible with a dynamical framework. This is because Peirce understood the interpretation of signs to be constituted by the production of signs (interpretant production), and because he also understood mental processes to be sign production in this same sense and not the locus of some intrinsically meaningful mental token. In this framework, there is no simple mapping between a sign vehicle (signifier) and what it refers to (signified). Rather, sign vehicles are physical forms that mark phases of a process of forms-modifying-the-production-of-other-forms. There is no final form that marks the terminus of the process. Indeed, for Peirce, what he calls the "final interpretant" is in effect a habit of interpretant generation, that is, a process organization.

To bring this semiotic analysis into alignment with dynamical theories it is necessary to understand signs as sources of constraint on dynamics and, as Pattee and Polanyi independently recognized, to understand that semiotic constraints must necessarily be distinct from the dynamics that they control. Both the code metaphor and the conception of cognition as mere correlated dynamics ignore this distinction. But a dynamical semiotic approach that treats sign vehicles (whether words or constraints on neural activity) as information structures that control the dynamics of the production of other sign vehicles can preserve the concept of representation, which satisfies the requirements specified by Brette (see also Bickhard 2009), without reducing it to a mapping relationship. It also provides a context for distinguishing modes of semiotic relations and semiotic differentiation processes in terms of different modes of constraint.

This reframing also illuminates the relationships of this debate to debates in information theory on the one hand and linguistic theory on the other.

Information theory, following the pioneering insights of Claude Shannon (1948), notoriously avoids any effort to deal with representational content or normative properties, such as accuracy and truth. And yet it provides a precise formalization for measuring the information content of a medium or a message within that medium, for optimally encoding a signal, and for compensating for noise (though both signal and noise are normative distinctions). Shannon's measure of information in a message is assessed in terms of the uncertainty that is thereby reduced by virtue of the constraint on its possible entropy. Thus, implicitly, it treats whatever semiotic value can be provided in a message as a function of this constraint.

Linguistics has also struggled with the code metaphor (as reflected in the famous signifier-signified relationship described by Ferdinand de Saussure [1959]), and critics of this conception have likewise attempted to reframe "languaging" in purely correlated dynamical terms. The assumption of a completely unstructured "arbitrary" correspondence between word sounds and meanings has led to a dilemma called the symbol grounding problem by Steven Harnad (1990b), on one hand, and has motivated theories of innate or culturally imposed grammatical principles and syntactic rules to explain the complex structure of languages, on the other. But a reframing of the linguistic structures in terms of constraints on dynamically grounded communicative processes also resolves these dilemmas (Deacon 2018; Rączaszek-Leonardi 2009; Rączaszek-Leonardi et al. 2018).

Acknowledgments. This work was supported by NCN OPUS 2018/29/B/HS1/00884.

Plasticity of the neural coding metaphor: An unnoticed rhetoric in scientific discourse

Giulia Frezza[a] and Pierluigi Zoccolotti[b]

[a]Metaphor Lab Amsterdam, University of Amsterdam, 1012 VB Amsterdam, The Netherlands and [b]Dipartimento di Psicologia, Sapienza Università di Roma, 00185, Rome, Italy.
g.frezza@uva.nl pierluigi.zoccolotti@uniroma1.it
https://www.researchgate.net/profile/Giulia_Frezza
https://www.researchgate.net/profile/Pierluigi_Zoccolotti?ev=hdr_xprf

doi:10.1017/S0140525X1900133X, e225

Abstract

The convincing argument that Brette makes for the neural coding metaphor as imposing one view of brain behavior can be further explained through discourse analysis. Instead of a unified view, we argue, the coding metaphor's plasticity, versatility, and robustness throughout time explain its success and conventionalization to the point that its rhetoric became overlooked.

Brette's thesis is that one main metaphor, that is, neural coding, has generated one main narrative, which carries a dominant standpoint on brain behavior that precludes other views, becoming almost undiscussed. By contrast, a scrutiny of the functioning of the coding metaphor shows a multipatterned usage of the metaphor and a broader metaphorical narrative. This plasticity of the coding metaphor clarifies why its pervasive and persuasive rhetoric spread across time almost invisibly.

Metaphors typically help us to understand complex ideas through more familiar elements by means of cross-domain mapping between "target" and "source" domains (Lakoff & Johnson 1980a). In "the brain is a computer," for instance, the target is the brain and the source is the computer; thus, we can imagine the brain in terms of elements of the computer (i.e., computing and programming). Both target and source domains are essential: they shape the metaphor and affect its meaning. Substituting "brain" with "mind" gives the metaphor "the mind is a computer," which changes its meaning; and replacing the source "computer" with "computational machine" gives "the brain is a computational machine," which was considered not metaphorical (Eliasmith 2003). The multiple targets of the "electric switch"

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

metaphor applied to epigenetics were considered a sign that the meaning of epigenetics is unclear and that it is a field in progress (Stelmach & Nerlich 2015).

In Brette's examples, instead of a unified view there are both different targets and sources of the coding metaphor. The targets vary, referring mainly to "treating information" (e.g., Somjen 1972, in Brette, sect. 1, numbered list, item 2); "perceptions" (e.g., Pouget et al. 2003, in Brette, sect. 2.1, para. 3); "representations" (e.g., Bickhard 2009, in Brette, sect. 3.1, para. 7); "neural activity" (Ashida & Carr 2011, in Brette, sect. 1, numbered list, item 1); and "brain activity" (Jazayeri & Movshon 2006, in Brette, sect. 1, numbered list, item 2). The source domain "coding" is associated with "encoding" (e.g., Brette, sect. 2.4, para. 6) and "decoding" (e.g., Brette, sect. 4.1, para. 1). A broader metaphorical narrative develops, which more generally targets the ideas of "reading" and "communication," connecting "coding" with "information" and with "encoding and decoding messages" to end with another metaphor about the brain "deciphering" hieroglyphs (Brette, sect. 5.2, last para.). This broad narrative, together with the varied targets and sources, indicates that the different voices in Brette's examples might stand for varied views of the coding metaphor instead of a unified one. This plasticity fulfills the well-known trait of metaphor in discourse, which needs to be "robust enough to carry certain implications from one context to another, but at the same time flexible enough to allow for different formulations in different contexts" (Hellsten 2003).

For example, the two targets of "neural activity" and "brain activity" allow for two different meanings. In most of Brette's examples of neural coding (e.g., sect. 4.1, para. 1, 2), the coding metaphor entails a dualistic perspective on brain functioning that duplicates the agents, e.g., "the brain," which makes decisions about stimuli deriving from "populations of neurons." Here the coding narrative presents a scenario in which the target "brain activity" stands for higher hierarchical functions (like "deciding," Brette, sect. 4.1, para. 1) than the target "neural activity." In this scenario, the neurons encode messages while the brain decodes them. In a different scenario, instead, the two targets are conflated because neural activity is a metonymy for brain activity (there are no "'little decoders' inside the head. … Rather, they are embedded in the synaptic weights between neighboring neurons" (Eliasmith 2003, p. 506). Here, the coding metaphor applies equally to both targets (i.e., brain and neurons). To get a clearer picture of the coding narrative, the metaphor should be specified according to its specific targets and sources in any given context so that the diversity of the underlying theoretical models, as well as the flexibility of the coding metaphor, can be better appreciated instead of concealed.

Moreover, the coding metaphor does not stand alone in the debate. At the core of the coding narrative there are two primary metaphorical models: "the brain is a computational machine" and the "information flow" of cognition (Dretske 1981; Searle 1980). For more than 60 years, these models outlined different views about cognition and brain functioning according to three main approaches: symbolicism, which focuses on a formal view of computation via symbols; connectionism, which emphasizes the parallelism of neural structures with networks; and dynamicism, which is an "environmentalist" view pointing to interactive and time-laden cognitive behaviors (Eliasmith 2003). Outlining Brette's criticism on neural coding against these pervasive cognitive metaphorical models frames his position within an older debate, that is, against a symbolicist and a computational view of the brain and in line with the dynamic approach.

We now turn to Brette's second point, that is, regarding the coding metaphor that establishes one dominant narrative of information that prevents imagining alternatives. Lakoff and Johnson (1980b) clarified the Gestalt mechanism of metaphor: by highlighting a given element (such as the computational view of the brain) other elements necessarily move to the foreground. In other terms, through metaphor the complexity of a phenomenon is restricted to show only one perspective at a time (Hellsten 2003). This better explains Brette's claim on the coding metaphor for which "the notion of information it carries seems to be the only possible one." Over the years, the narrative developed by the metaphor became conventionalized and established "self-evident ways of seeing things, even to the degree where no alternatives are imagined" (Hellsten 2003). It is the plasticity of the coding metaphor, that is, its robustness through change, that explains why for more than 60 years it spread and served across time and debates, allowing for its conventionalization, similarly to "brain-computer" and "information" metaphors. The pervasive and persuasive effects of the metaphorical narrative hinder the fundamental self-correcting trait of science that aims to provide counterexamples of dominant theories instead of just supporting them. This is part of the well-known circuit of metaphor in scientific discourse that entails potential risks, from misunderstanding to hype in communication, calling for more responsibility in language use, which should reinforce our vigilance about metaphor and metaphorical narratives in science (Ball 2011; McLeod & Nerlich 2018).

Our understanding of neural codes rests on Shannon's foundations

Charles R. Gallistel

Rutgers Center for Cognitive Science, Rutgers University, Piscataway, NJ 08854.
galliste@ruccs.rutgers.edu
https://ruccs.rutgers.edu/gallistel-research-interests

doi:10.1017/S0140525X19001249, e226

Abstract

Shannon's theory lays the foundation for understanding the flow of information from world into brain: There must be a set of possible messages. Brain structure determines what they are. Many messages convey quantitative facts (distances, directions, durations, etc.). It is impossible to consider how neural tissue processes these numbers without first considering how it encodes them.

Brette's treatment of information theory does not do justice to the role of the receiver in Shannon's (1948) theory. The three elements of communication essential to understanding the role of codes in neural communication are not "correspondence, representation, causality"; they are source, signal, and receiver. The *receiver* must have a *set of possible messages* it may receive about *some state of the world* (the source) by way of a *signal*. It must also have a *probability distribution* over the set of possible messages, because Shannon's formula makes information a property of probability distributions. Absent a distribution, there is no

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

measure of information. Shannon's theory establishes the conceptual foundation for a scientific understanding of the flow of information from the world into and within brains (Gallistel, in press).

Brett fails to ask the first question that must be asked when applying Shannon's theory to understanding world-brain communication: What determines the set of possible messages? Not the world; its role is passive. Brain structure determines the sets of possible messages; they are all and only the sets of messages its highly differentiated structures enable it to receive. Sensory transducers and the signal processing machinery that extracts information about distal stimuli (things out there in the world) from the signals generated in those transducers by proximal stimuli (the stimuli that act directly on the transducers) determine the messages a brain can receive.

Color vision provides an illustrative example. The set of messages the human brain receives about the reflectance spectra of surfaces is determined by the distinguishable locations in a neural vector space with three bipolar dimensions. Neither the dimensionality of the vector space nor the bipolarity of the vectors that encode color is a property of spectra. The brain imposes this encoding when it creates the three types of cones in the retina and the multiple stages of signal processing that map from cone photon catches to the encodings that mediate color percepts. In setting up three and only three cone types, brain epigenesis establishes the dimensionality of the space. In setting up the circuitry for subtracting the signal from one cone type from the signal of another cone type, it establishes the bipolarity of the representational structure. The bipolarity creates a distinctive feature of color perception, the mutual exclusivity of certain color pairs (red-green and yellow-blue, for example).

Shannon's source-coding theorem enables us to understand why the brain imposes this structure on the color messages it receives: An efficient code must reflect the source statistics. In the case of color, the source statistics are the statistics of the reflectance spectra of surfaces in the natural world. In principle (and in a lab equipped with monochromators), the intensity of light at any wavelength is independent of its intensity at any other, but in the world, reflectance spectra have massive redundancies, because the intensity at one wavelength is highly predictive of the intensities at neighboring wavelengths. This redundancy greatly reduces the available information. The brain's way of encoding color captures a large part of the information available from the reflectance profiles of surfaces in the natural world (Boker 1997; Maloney 2003). Vague talk about the "dynamic, circular, distributed" nature of brain processes does not deliver this kind of insight.

A computing machine like the brain has four material foundations: its signals, which transmit information from place to place within the machine: the symbols in its memory, which transmit information from the past into the future; the machinery for executing signal processing operations; and the machinery for executing operations on symbols (Gallistel & King 2010). The machinery for processing the signals and operating on the symbols cannot be designed – if one is building the machine – or understood – if one is reverse engineering it – until one has decided on, or come to know, the code or codes by which the information will be, or is, represented in the signals and the symbols. The foundational role of the code is clear to computer engineers and to those who know the history of molecular biology (Judson 1980). Its importance to neuroscience is well illustrated by the Rieke et al. (1997) book, *Spikes: Explorations of the Neural Code,* which Brette cites, but otherwise ignores.

Much of the information conveyed by neural signals and stored in neural memory is quantitative. The physically realized representatives of quantities in a computing machine (e.g., bit patterns) are what computers scientists understand by numbers. The brain performs arithmetic operations on the signals and symbols, which is one good reason for conceptualizing brain function in computational terms.

An example of neural arithmetic is the time-compensated sun compass. Animals learn and store in memory the sun's azimuth as a function of the time of day (Dyer & Dickinson 1996; von Frisch 1967; von Frisch & Lindauer 1954). They can then steer by the sun while flying a compass bearing to a food source (whose location may have been obtained the previous day by following the dance of another forager [Menzel et al. 2011]). To enable that behavior, their brain must subtract the current solar azimuth from the desired compass course to obtain the current solar bearing of the source, the angle at which they must hold the sun's image while flying to their destination.

Understanding the neural machinery that performs angular subtraction requires understanding how the brain encodes angular quantity (Gallistel 2018). Proponents of dynamical system theory do not seem prepared to consider how the brain does the arithmetic required in navigating. This refusal frees them from the need to ponder how it encodes quantities (Gallistel 2017; 2018).

An understanding of the brain's codes is as essential to neuroscience as an understanding of the genetic code is to biology. Shannon's theory is now, and is likely to remain, the foundation of that understanding (cf. Qian & Zhang 2019; Stevens 2018).

The origin of the coding metaphor in neuroscience

Justin Garson

Department of Philosophy, Hunter College of the City University of New York, New York, NY 10065.
jgarson@hunter.cuny.edu www.justingarson.com

doi:10.1017/S0140525X19001316, e227

Abstract

To assess Brette's proposal to expunge "coding" from the neuroscientist's lexicon, we must consider its origins. The coding metaphor is due largely to British nerve physiologist Edgar Adrian. I suggest two ways that the coding metaphor fueled his research. I conclude that the debate today should not be about the "truth" of the metaphor but about its continuing utility.

Brette, in his provocative article, gives a number of arguments for his proposal that we expunge the coding metaphor from the neuroscientist's lexicon. To properly evaluate his proposal, however, we should consider the metaphor's *origins*. When did neuroscientists begin describing the neural response to a stimulus as a coded message? Why did they begin using that metaphor? What benefit, if any, did they derive from using the metaphor?

I have investigated these questions extensively, and concluded that the British nerve physiologist Edgar Douglas Adrian (1889–1977) deserves most of the credit for introducing the metaphor

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

(Garson 2015). Adrian's most notable accomplishment was his recording, in 1925, of the action potential of a single sensory neuron (Adrian and Zotterman 1926). This was the achievement for which he was awarded the Nobel Prize in Physiology and Medicine in 1932. Additionally, he formulated what he understood to be the basic laws governing the action potential: the all-or-none principle, rate coding, and adaptation (see Rieke et al. 1997, pp. 3–8). Following his achievement, he spent the next decade attempting to demonstrate the universality of these basic laws across different neuron types and in different species.

The year 1925 marked another turning point in Adrian's work. From that time onward, his work was permeated with coding metaphors. He described the neural response in terms of the transmission of *messages*, *signals*, and *codes*. In a revealing analogy, he noted that these messages are "scarcely more complex than a succession of dots in the Morse Code" (Adrian 1932, p. 12). In his use of linguistic metaphors he distinguished himself from his contemporaries, such as Charles Sherrington, Herbert Gasser, and Alan Hodgkin, who, at the time, preferred the colorless language of "impulses," "reactions," "activity," and "disturbances" to describe the neural response.

This is not to say that nobody before Adrian had described the nervous system as a communication device. That analogy dates back to the middle of the nineteenth century, when pioneers like Hermann Helmholtz and Emil du Bois-Reymond, in their popular writings, compared the nerves to a telegraph system that shuttles news and instructions from the body to the brain and back. Adrian's use, however, was quite distinctive. Unlike his predecessors, he attributed a highly specific, language-like code to the neuron. For Adrian, neural responses were, as Brette puts it, "hieroglyphs to be deciphered" (last line of text).

We can speculate on the historical, sociological, and technological context that might have prompted Adrian to think about the neuron as a coding device. I have argued that the widespread military use of radio communication during World War I played a role in this terminological innovation. But that is hardly the issue here. The issue here is this: Did the coding metaphor actually benefit Adrian's research? Did it meaningfully advance neuroscience? Was it, in its time, a scientifically fruitful metaphor?

I think the answer to these questions is a decisive "yes." The metaphor let Adrian formulate and test questions that had never been systematically posed before. There were at least two fields of investigation that the coding metaphor opened for him. First, it allowed Adrian to shift his attention away from the mechanistic details of the action potential (e.g., how the impulse propagates through narcotized nerve) and toward the *abstract correspondences* between the pattern of sensory stimuli and the neuron's patterned response (e.g., how a rapidly changing stimulus modulates the neural response). By posing questions about these abstract correspondences, Adrian was able to gather evidence for what later became known as "rate coding": For some sensory neurons, spike frequency approximates an exponential function of the intensity of the stimulus. It is hard to see how one would even formulate such questions without using the coding metaphor.

Second, the metaphor allowed him to pose questions about the *purpose* or *end* of various "coding" schemes. Put differently, it allowed him to reason teleologically about the brain. Why does the brain use rate coding, rather than some other coding principle, to represent rapidly changing stimuli? Consider, for example, the principle of adaptation, which describes how some sensory neurons eventually stop responding to an unchanging stimulus. For Adrian, this principle could be explained teleologically, as a bulwark against the pointless, and metabolically costly, production of redundant messages (Adrian 1928, p. 99). Some biologists bristle at the mention of "teleology," but the simple fact is that teleology cannot be eliminated from biology (Garson 2019). Every time we ask a question about the *function* of a trait (what is the function of zebra stripes?) we are engaged in teleological reasoning. Moreover, such reasoning is usually harmless, as teleological questions can often be restated as respectable evolutionary questions (why did zebra stripes evolve?), rather than questions about the intentions of a divine being.

I have discussed the origins of the coding metaphor, but how does this bear on Brette's proposal to expunge "coding" from neuroscience? For everything I have said here, Brette might still be correct that the coding metaphor is otiose. He might insist that, while the metaphor might have been useful in the early days of the neurosciences, it is no longer so. Because I am a philosopher and not a neuroscientist, I am hardly in a position to survey the current state of the science and make a pronouncement of such scope. Still, thinking about the origins of the coding metaphor can helpfully frame what I expect to be an ongoing and lively debate. It seems to me that the question of eliminating "coding" should not be a referendum on the *truth* of the coding metaphor, but on its *utility*. In other words, in assessing Brette's proposal, we should not get bogged down wondering whether the brain "really" encodes information about the world. It does not. It is a metaphor; like all metaphors, it gives us a partial and imperfect picture of what the brain is doing. The questions, rather, are these: Is the metaphor still useful for us, today? Does the usefulness of the metaphor outweigh its inaccurate connotations? Or, has it outlived its usefulness entirely?

From mental representations to neural codes: A multilevel approach

Jon Gauthier[a], João Loula[a], Eli Pollock[a], Tyler Brooke Wilson[b] and Catherine Wong[a]

[a]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139 and [b]Department of Philosophy, Massachusetts Institute of Technology, Cambridge, MA 02139.
jon@gauthiers.net jloula@mit.edu epollock@mit.edu tbw@mit.edu catwong@mit.edu
http://foldl.me https://joaoloula.github.io/ https://www.elibpollock.com/ https://sites.google.com/site/tylerbrookewilson/ https://web.mit.edu/zyzzyva/www/

doi:10.1017/S0140525X19001390, e228

Abstract

Representation and computation are the best tools we have for explaining intelligent behavior. In our program, we explore the space of representations present in the mind by constraining them to explain data at multiple levels of analysis, from behavioral patterns to neural activity. We argue that this integrated program assuages Brette's worries about the study of the neural code.

We advocate an approach that grounds claims about neural codes in evidence about the brain and mind. Under this approach, the

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

search for neural representations begins with an understanding of the *task* that an organism solves, in the spirit of rational analysis (Anderson 1989). The next step is to propose computational models capable of solving this task. These models constitute candidate hypotheses for the representations and computations employed within the brain, allowing us to establish principled constraints on possible neural codes and give strict satisfaction conditions for their implementation.

This program portrays brains as *representational* and *computational* devices. Brette's arguments challenge this foundational idea. We address his arguments on three fronts: (1) the indeterminacy of claims about the neural code, (2) the ability of neural networks and dynamical systems to instantiate structured representations, and (3) the historical success of the search for representations in the brain.

First, Brette argues that many studies of neural codes fail to account for relevant contextual features in ecological behaviors. Brette criticizes studies of neural codes driving sound localization, for example, which neglect how contextual features such as sound amplitude interact with frequency. Brette draws the conclusion that neuroscientific evidence underconstrains the search for the contents of neural codes. But this search can be helpfully constrained by the computational theories introduced above. As accounts of ecological behavior, computational models both generate and constrain our hypotheses about the environmental features relevant for a given task. Such research puts us in a position to make educated guesses as to the environmental features that are likely to be represented in neural codes.

For example, computational models of vision and audition describe how ecological tasks (e.g., object recognition) are carried out via computational mechanisms. These models demonstrate both *why* representations with particular content are present in the mind and *how* they are used to produce behavior. After being empirically tested against observed organism behavior, their internal representations can be used to guide the search for neural codes (see, e.g., Rajalingham et al. 2018; Młynarski & McDermott 2018). This constraint – that neural codes must implement the empirically validated representations of computational models – greatly reduces an otherwise open-ended and intractable search for neural representations.

Second, computational-level models can help us narrow down candidate neural representations, but such a search will be fruitful only if representations can actually be realized in neural circuitry. Brette argues that this cannot be done. We think that the evidence suggests otherwise: Both connectionist models and neural dynamical systems are capable of implementing structured representations.

Brette worries that neural codes composed of cell assemblies can encode only "objects or features to be related, but not the relations between them," and points to connectionist models as evidence. But connectionist researchers have worked to address this issue and have adapted models to learn relational structure, since almost the inception of the field (McClelland 2003; Smolensky & Legendre 2006). Indeed, modern neural networks continue to challenge and broaden our understanding of the kinds of relational structures they *can* learn to represent when designed and trained carefully: from abstract, hierarchical syntactic dependencies in natural language (Gulordava et al. 2018), to dynamics in intuitive physical systems (Fragkiadaki 2016; Chang 2017), to complex relationships between objects in visual scenes (Hudson and Manning 2018; Johnson 2017; Yi 2018). The need to represent complex structures can and should drive us to think creatively about how cell assemblies can instantiate them, rather than abandon the project because of arguments from inconceivability leveled by its detractors.

Brette also alleges that "dynamical systems cannot in general be mapped to algorithmic descriptions" (sect. 4.1, para. 1). But recent theoretical work offers paradigms using recurrent networks to do precisely that (Eliasmith and Anderson 2004; Mastrogiuseppe and Ostojic 2018). Concurrently, additional work has suggested that it is possible to map latent states of behavior to dimensions of neural activation (Afraz and Jazayeri 2017) and verified predictions about the encoding of more complex tasks in higher-dimensional spaces (Gao et al. 2017). Once we accept that population-level neural dynamics provides a substrate for representations, it makes sense to look for algorithmic transformations of those representations, that is, a code. All signs indicate that a major goal of neuroscience in the near future will be identifying context-invariant subspaces of neural activity that act as such a code (Saxena and Cunningham 2019).

Finally, this integrated approach – one that begins by understanding the task, proposes algorithms to compute it, and finds its representations in the brain – is exemplified by research on temporal difference learning. Temporal difference is a reinforcement learning algorithm that requires representing the current reward and the predictive value of the current state, and computing reward prediction error. Early research on the neural implementation of reinforcement learning showed that reward prediction error correlated with phasic dopamine signals (Montague 1996; Schulz 1997). Correlational evidence, as Brette is quick to point out, does not imply neural coding. But research on temporal difference learning in the brain has gone far beyond correlation. This single representational model has yielded continuous empirical successes: predicting neural responses, both quantitatively and qualitatively (Niv 2009), causally manipulating neural responses and observing predicted behaviors (Steinberg 2013), and discovering the mechanisms responsible for key symptoms of Parkinson's disease (Frank 2004). If such representations are not actually instantiated in the brain, this streak of results spanning more than two decades would be nothing short of a miracle.

We believe that principled constraints can be placed on the search for neural codes. These constraints come from integrating multiple levels of analysis, including an understanding of the task being solved, a hypothesis space of algorithms capable of solving it, and behavioral and neuroscientific evidence to decide between candidate hypotheses. We've highlighted work showing how structured representations can be implemented in connectionist and dynamical systems models. We've described one strikingly successful search for neural codes in the brain. Together, these successes suggest that Brette's skepticism is unfounded. This integrated search for neural codes remains the best framework for understanding the brain.

A clash of *Umwelts*: Anthropomorphism in behavioral neuroscience

Alex Gomez-Marin

Behavior of Organisms Laboratory, Instituto de Neurociencias (CSIC-UMH), 03550 San Juan Alicante, Spain
agomezmarin@gmail.com https://behavior-of-organisms.org/

doi:10.1017/S0140525X19001237, e229

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

Abstract

Brains enjoy a bodily life. Therefore animals are subjects with a point of view. Yet, coding betrays an anthropomorphic bias: we can, therefore they must. Here I propose a reformulation of Brette's question that emphasizes organismic perception, cautioning for misinterpretations based on external ideal-observer accounts. Theoretical ethology allows computational neuroscience to understand brains from the perspective of their owners.

An apparently innocuous word in Brette's question is a major source of confusion but also contains a great deal of the answer. Is coding a relevant metaphor for "the" brain? Yes and no. It depends on whose brain we are talking about. For the scientist studying the animal, coding is certainly relevant (at least, as the ubiquity of such figure of speech attests in current neuroscience). But, insofar as we are interested in the animal and its brain, the answer is likely no. The mantra "stimulate, record, correlate" misses the point of the organism. It is *for us, by us*. That the experimenter's model can decode the signal does not mean that the brain can or does. The information necessary to make sense of the data in terms of coding is seldom available to the organism, upon which coding is predicated. This creates a can-ought problem: a description of what the neuroscientist can do prescribes what the animal must do. Such implicit tension pervades most of the disagreements that Brette's question shall spur. The problem, I believe, is deeper than coding: There is a conflict of interests between the scientist and the laboratory animal.

Biology is the science of living beings. Organisms are centers of action. As such, perspective matters. To be an organism is to have a point of view. All animals share a common world but not all animals have a world in common. Each living organism has its own *Umwelt* (meaningful environment), which is different than its *Umgebung* (physical surroundings): A tree is a tree, but a tree for an ant has little to do with a tree for a carpenter (Uexküll 1926). What is meaningful for an organism – or even what is possibly apprehensible – need not be meaningful for the scientist studying it, and *vice versa* (a concrete and pervasive example: stimuli are more the experimenter's output than the animal's input). The use of the definite article ("*the* brain") or the indefinite pronoun ("*one* finds") is so delicate in biology. It easily blurs the subject (I? you? the mouse? what mouse?), unbinding grave conflations and misleading thought and interpretation. Eloquently said, "Hedgehogs as such do not cross roads (…). On the contrary, it is man-made roads that cross the hedgehog's millieu" (Canguilhem 2008, p. 22). Rather than being an exception, coding illustrates such misattribution. Paraphrasing, we could say that cat brains as such do not encode stripes, but it is stripes that we decode from the cat's brain. A clash of *Umwelts* (*Umwelten*, in proper German) is going on in our laboratories.

The notion of *Umwelt* has no place in physics; it does not violate physics, but it is not reducible to physics either. Living beings inhabit a world of meaning that includes but exceeds the physical world of masses and forces, and even more so the mathematical world of zeros and ones. The appreciation of the uniqueness of biology discords with a cornerstone of the scientific approach: objectivity. Of course we always observe reality from a viewpoint, explicitly or implicitly chosen. But it is ultimately deemed irrelevant. Objectivity, then, is the pretense of self-exclusion from the phenomenon under study. The observer vanishes in classical physics (also in biology). By means of a representation of things that ultimately does not depend on the reference system, an observer-independent reality is erected. Yet, "[o]n the strength of the immediate testimony of our bodies *we* are able to say what no disembodied onlooker would have a cause for saying: (…) the point of life itself: its being self-centered individuality" (Jonas 2001, p. 79). From subjectivity we have prodigiously built an objectivity that can dispense with the former. However, upon inspection, objectivity becomes a particular kind of intersubjective consensus. This is biology's scotoma: We are subjects whose objects of study are subjects too.

In behavioral neuroscience there is an observer-observed gap. Physiology aspires to study the inner workings (brain) of an organism from the outside (scientist's perspective); ethology strives to understand the outer happenings (behavior) from the inside (animal's perspective). Isn't the neurophysiologist's decentering a covert self-centering? Sticking electrodes is not sufficient to know what it is like to be a rat. But, how to look through the animal's eyes? A cute example is *Turtle Geometry*: it actually matters if a turtle traces a circle by solving the $x^2 + y^2 = r^2$ equation, or by iterating a "run and turn" procedure. Both are mathematically equivalent (from an external ideal observer, perhaps indistinguishable, even irrelevant) but biologically they are not the same. There is much to gain from discovering "the range of complicated things a turtle can do in terms of the simplest things it knows" (diSessa & Abelson 1981, p. 3). What is it to make sense from the animal's perspective when it does not do so the way we do? Such is the paradox: The *Umgebung*, the objective world of scientists, can be part of our human *Umwelt* (we do not feel neutrinos crossing our bodies, but we can detect them in bubble chambers), but it collides with the *Umwelt* of the animal, which is never an *Umgebung*. Neuroscientists yearn for neural codes; the animal has no clue.

Neuro-ethology is actually meta-engineering: our problem is to solve how animals solve their problems – to scientifically empathize with each creature. This entails a revision of Bernard's (1957, p. 103) foundational words: The scientist "no longer hears the cry of animals, he no longer sees the blood that flows, he sees only his idea and perceives only organisms concealing problems which he intends to solve." By reformulating Brette's question, my intention here has been to emphasize that computational neuroscience can benefit from the insights of theoretical ethology to transform its anthropomorphic bias. To crack codes, "it would suffice that we be angels. But to do biology, even with the aid of intelligence, we sometimes need to feel like beasts ourselves" (Canguilhem 2008, p. xx). The question then is not so much whether coding is relevant or wrong, but to what extent it is misleading. We must then ask: Whose brain is the coding metaphor relevant for?

Acknowledgments. I thank Ehud Ahissar, Konrad Kording, Spyridon Koutroufinis, Ibrahim Tastekin, Zach Mainen, and specially Asif Ghazanfar for insightful discussions.

Beyond metaphors and semantics: A framework for causal inference in neuroscience

Roberto A. Gulli

Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027; Center for Theoretical Neuroscience, Columbia University,

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

New York, NY 10027; Department of Neuroscience, Columbia University, New York, NY 10027
r.gulli@columbia.edu
http://robertogulli.com

doi:10.1017/S0140525X19001389, e230

Abstract

The long-enduring coding metaphor is deemed problematic because it imbues correlational evidence with causal power. In neuroscience, most research is correlational or conditionally correlational; this research, in aggregate, informs causal inference. Rather than prescribing semantics used in correlational studies, it would be useful for neuroscientists to focus on a constructive syntax to guide principled causal inference.

In his article, Brette argues that the "coding metaphor" in neuroscience is inappropriate and misleading because it leads to false interpretations of causality. Brette states that "by postulating that neural codes are representations, we imply that these codes have a causal impact on the brain" (sect. 4.2, para. 1). However, this is implausible since "[in the] technical sense … the word *code* is used as a synonym for correlate" (sect. 1, para. 4). Restated, the coding metaphor is problematic because it can imply causal function where sufficient evidence to support causal inference does not exist. By relying on this criticism, Brette commits to a broader error: He interprets that isolated correlations, conditional correlations, and statistical inferences between neural activity and function support or refute causal inference. Isolated pieces of correlational or statistical evidence are insufficient to demonstrate a causal relationship between neural activity and functions, perceptions, or behaviors, and should be considered in aggregate to form the basis of causal inference. For this reason, it would be helpful for those seeking to design and interpret experiments to adopt a constructive framework for causal inference in neuroscience.

The correlational nature of individual studies in neuroscience has been explicit since the dawn of electrophysiology, when Caton (1875) stated that "[t]he electric currents of grey matter appear to have a relation to its functions." Contemporary studies of neural activity and function are still strictly correlational, despite advances in recording and analysis methods. Traditional statistical techniques are agnostic to causal relationships between variables and thus cannot determine causality (Pearl et al. 2016). Experimental interventions that support causal inferences between brain (dys-)function and behavior have long been sought (Dodds 1878; Ferrier 1886). However, even studies that use modern versions of these "causal" techniques (optogenetic, chemogenetic, electrical, and pharmacological modulation) provide correlations conditioned on perturbation. Causal inferences on the basis of single experimental results should be tempered because of plausible confounding and off-target effects (Jazayeri & Afraz 2017).

Insights from other fields provide a clear path toward causal inference with individually circumstantial pieces of evidence. The most influential perspective may be that of medical statistician Austin Bradford Hill, who described nine "viewpoints" that guide causal inference in epidemiology when randomized controlled trials are not possible (Hill 1965; see also Phillips & Goodman 2004). Here these viewpoints are adapted to form a Bradford Hill-inspired framework for causal inference in neuroscience, where aggregated observational and interventional studies support causal inference:

1. **Correlational evidence:** Relationships between measurements of neural activity and experimenter-defined responses (whether in downstream neural activity, other physiological or behavioral outcomes). These relationships can be characterized through a variety of forward and backward modeling techniques (see, e.g., Anderson 2019; Baayen et al. 2008; Marinescu et al. 2018; Rougier 2019; Saxena and Cunningham 2019; Song et al. 2013; Wang and Yang 2016).
    i. *Strength:* Does the neural activity explain a reasonable amount of variability in the response?
    ii. *Consistency:* Does the neural activity reliably produce the outcome?
    iii. *Specificity:* Is the observed relationship between neural activity response unique or one of a vast array of potentially confounding correlations?
    iv. *Relationship curve:* Is there a clear geometric relationship between neural activity and the response?
    v. *Temporality:* Does the neural activity consistently precede the response in time?
    vi. *Mechanistic plausibility:* Is there a plausible mechanism whereby neural activity may produce response?

2. **Conditionally correlational evidence:** The effect of direct or indirect modulation of neural activity on experimenter-defined outcomes. Modulation includes loss-of-function and gain-of-function "causal" manipulations that are under control of the experimenter.
    i. *Strength:* Does modulation of neural activity explain a reasonable amount of variability in the response?
    ii. *Consistency:* Does modulation of neural activity reliably produce the predicted outcome?
    iii. *Specificity:* Does modulation of neural activity lead to a prescribed outcome or one of a vast array of potential effects?
    iv. *Relationship curve:* Is there a predictable and replicable geometric relationship between modulation of neural activity and the response?
    v. *Temporality:* Does the predicted effect follow the perturbed neural activity at a reasonable delay?
    vi. *Coherence:* Is the predicted effect of modulation of neural activity coherent with other strong hypotheses?
    vii. *Analogy:* Does a modulation of closely related neural activity patterns produce similar effects?

With this framework in mind, one should reconsider Brette's claims related to neural codes and causal inference. For example, Brette states that "BOLD (blood oxygen level-dependent) signal … encodes visual signals in the same technical sense that the firing of neurons encodes visual signals" (sect. 4.2, para. 2). Functional magnetic resonance imaging and electrophysiology studies are both correlational, but Brette's assertion is deeply flawed in important ways. In fMRI and electrophysiology, fundamentally different biological activity is associated with stimulus or behavioral response of interest (Goense and Logothetis 2008). Thus, the mechanistic plausibility of a link between neural activity and the experimental condition differs. Furthermore, the spatial specificity and temporality of visually evoked activity cannot be similarly addressed across techniques (Sejnowski et al. 2014). These factors are critically important in guiding causal inference, and therefore, each technique uniquely contributes toward causal inference. To suggest that BOLD signals and action potentials encode visual stimuli in the same technical sense is a conspicuous oversimplification. In this example, the

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

proposed framework for causal inference aids in articulating the relative strengths and weaknesses of different experimental approaches. Furthermore, it provides guidelines for making causal inferences by aggregating individual pieces of evidence that are insufficient in isolation.

Regarding the causal relationship between the physical world and thought, Haugeland (1985, p. 106) stated, "If you take care of the syntax, the semantics will take care of itself." This axiom presents a useful analogy: with a proper framework to describe the syntax (rules and criteria) of causal inference in neuroscience, Brette's claim – the coding metaphor perpetuates inappropriate causal inference – is reduced to an innocuous semantic debate. His further claim that metaphors perpetuate "semantic drift" (sects. 2.1, 2.2, and 3.1) should be addressed not by further semantic prescriptions, but by adhering to reasoned syntax. These semantic debates distract from the ultimate goal of discovering robust, causal relationships between the many levels of organization in the brain and behavior.

Acknowledgments. I thank David L. Barack, Matthew L. Leavitt, and Rishi Rajalingham for criticism and comments.

Codes, communication and cognition

Stevan Harnad

Department of Psychology, Université du Québec à Montréal, Montréal (Québec) H3C 3P8, Canada; Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1TW, UK.
harnad@soton.ac.uk

doi:10.1017/S0140525X19001481, e231

Abstract

Brette criticizes the notion of neural coding because it seems to entail that neural signals need to "decoded" by or for some receiver in the head. If that were so, then neural coding would indeed be homuncular (Brette calls it "dualistic"), requiring an entity to decipher the code. But I think Brette's plea to think instead in terms of complex, interactive causal throughput is preaching to the converted. Turing (not Shannon) has already shown the way. In any case, the metaphor of neural coding has little to do with the symbol grounding problem.

Both Shannon's (1948) *information* and Turing's (1936) *computation* are important in cognitive science. Shannon is concerned with the faithfulness of signal transmission in communication, and Turing is concerned with what algorithms can do. Cognitive science is concerned with what organisms (hence their brains) can *do*, and *how*.

Cells (including neurons) transmit signals. This is already true in plants (Baluska & Mancuso 2009) and of course also in machines. And organisms certainly do things. Which of the things organisms do are "cognitive" and which are "vegetative" is mostly just a definitional matter, but it is probably overstretching the notion to say that paramecia or hearts are "cognizing." The examples are nevertheless instructive for cognitive science, because paramecia, hearts, and organisms with brains are all systems that can *do* things. So are computers and robots, for that matter. Hence finding a causal explanation of how one of them does what it does may provide useful lessons for explaining the others.

Let's start with the heart, an example used by Brette. What does the heart do? It pumps blood. No metaphors. The heart literally pumps blood, and cardiac science has successfully reverse-engineered the heart (to a close approximation). We know how the heart does it – and part of the proof that we know how is that we can apply and test our hypotheses about how the heart pumps blood by building a synthetic model of a heart, plugging it into the heart's inputs and outputs, and testing whether it can pump blood. If it can, the artificial heart passes the "Turing Test" for cardiac function.

So what does the (human) brain (and body) pump? Human behavior. Or, rather, human behavioral *capacity*. What people *can do*. Let's forget about what portion of that capacity counts as cognitive and what proportion is just vegetative (like cardiac function): It all consists of the capacity of a (living) system to do certain things. Now the challenge is to explain how.

Turing (1950) provided the ground rules: You have an explanation if you can design a system that can do everything a human being can do, indistinguishably – *to* a human – *from* a human. If your interest is just in "cognitive" capacities, then just generate those, ignoring the vegetative capacities (or at least those that are not essential for generating the cognitive capacities). Cognition, like Justice Potter Stewart's pornography, may be hard to define, but we know it when we see it. And the capacity to interact with the dynamic world of objects and events and their properties (including words describing those objects, events, and properties) indistinguishably from the way humans do is surely cognitive, if anything is.

There is one more thing: Humans don't just do: They also feel. It *feels like something,* to a human, to be seeing and doing what humans can see and do. But the capacity to feel eludes Turing's program for cognitive science. It's something our brains pump invisibly. Turing (1950) accordingly brackets it. But it keeps making disruptive peekaboo appearances in our attempts to reverse-engineer cognition, as we shall see.

One of the main hypotheses about how the brain pumps cognitive capacity is via computation, Turing computation. Computation is the manipulation of "symbols" (arbitrary formal objects) on the basis of rules operating only on the symbols' shapes ("syntax"), not their meanings ("semantics"), to generate certain symbolic outputs from certain symbolic inputs. That's what algorithms do. (An intuitive example is the rule we all learned in school for extracting the roots of quadratic questions: "minus *b* plus or minus the square root of ….")

Algorithms are like recipes: apply them to the symbolic ingredients and you can explain how to bake a symbolic cake. Computation is very powerful; just about everything in the universe can be encoded symbolically and explained computationally, including cardiac function. The right algorithm can pump symbolic blood. And you can show that the algorithm really works by applying it to build a synthetic heart that really passes the cardiac Turing Test (TT) and pumps blood. But to do that, you have to "interpret" the symbolic code and implement it in material form, just as a formal recipe for a cake needs to be implemented in material form, using the real ingredients referred to by the symbols, to generate a real cake.

So, despite its enormous power, computation cannot be all there is to cognition. Searle (1980) showed, famously (in this journal), that a computer is not cognizing even if it can pass the TT because Searle too could pass the Chinese TT by executing the symbolic code without understanding a word of Chinese. Why can't he understand? Because there is no connection between the symbols in the code and the objects in the world that they are interpretable

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

as being about. Interpretable by whom? The user or the executor of the code. But the meaning itself is not in the code.

That is the symbol grounding problem (Harnad 2006). Simple solution: The TT must not be merely symbolic (verbal). It must test not only what the candidate can say, but also all the other things a human cognizer can do in interacting with the objects in the world that the verbal TT is merely chatting about. The candidate has to be a robot (Harnad 2009). And a Turing robot is not just a computer, manipulating formal symbols; it is a dynamical system, able to interact with the objects in the world. Its symbols are grounded in its capacity to identify and interact with their referents indistinguishably from the way we do.

Now to neural "codes": Brette is right that it would be homuncular (although he calls it "dualistic") to think of input to sensory receptors – activity along sensory pathways to sensory and sensorimotor regions in the brain and then onward to motor regions and pathways to motor effectors – as encoded signals being transmitted in order to be decoded by a receiver, as in telegraphic communication of Morse code from a sender to a receiver. There is no homunculus on the receiving end. It's all just a dynamic causal process constituting the organism's capacity to do what it can do, some of it output in response to immediate sensory input, some of it generated by endogenous processes.

But it is harmless to call the neural activity along sensory input pathways a "neural code." Shannon's communication theory is about the end-to-end fidelity of signal transmission (of analog or digital signals); it is not about cryptography, let alone about the interpretation of computational algorithms or of natural language. To show that there is a substantive issue involved here, Brette would have to show that there is a nontrivial chunk of performance capacity (even the detection of interaural time differences) that cannot be explained causally if we insist on calling the activity occurring along the sensory input pathways a "neural code." (Brette's preferred notion of "neural representations," by the way, sounds just as homuncular to me as the idea of neural codes: "representation of what, to whom?" Ditto for "internal model.")

Let me close with Brette's fleeting mention of "percepts." This is an instance of the "peekaboo" influence of homuncular thinking. Psychophysics, too, can only study what the organism does (input/output), not whether or how it feels like something to do it. Sensorimotor activity is only perceptual if it is felt. I don't doubt that it feels like something to detect an interaural time difference, just as it feels like something to understand Chinese. But although symbol grounding and Turing-testing may be "easy" (in principle, if not in practice), explaining how and why organisms *feel* rather than just *do* is and remains notoriously hard.

Neural code: Another breach in the wall?

Chloé Huetz, Samira Souffi, Victor Adenis and Jean-Marc Edeline

Institute of Neuroscience, NeuroPSI, UMR CNRS 9197, Université Paris-Sud, 91405 Orsay, France.
chloe.huetz@u-psud.fr samira.souffi@u-psud.fr victor.adenis@u-psud.fr
jean-marc.edeline@u-psud.fr http://neuro-psi.cnrs.fr/

doi:10.1017/S0140525X19001328, e232

Abstract

Brette presents arguments that query the existence of the neural code. However, he has neglected certain evidence that could be viewed as proof that a neural code operates in the brain. Albeit these proofs show a link between neural activity and cognition, we discuss why they fail to demonstrate the existence of an invariant neural code.

By questioning the existence of the neural code, Romain Brette opens again a strong debate between representational views of the brain (cognitivism and computationalism) and sensorimotor/enaction theories (O'Regan and Noë 2001; Varela et al. 1991), his preference being the latter. According to his view, all cognitive functions, particularly action and perception, are viewed as means to interact with the world, without the need to build internal representations of it. Neural activity during perception should be viewed as the result of the organism's interaction with the world, taking into account all possible influences, such as its internal state and its actions resulting in a given percept. Therefore, as the brain does not manipulate representations, it is senseless to try to decipher any code supposed to encrypt representations in neural activity. The results of three research fields focusing on proving that a particular neural code is at play should be addressed by Brette's review to strengthen his point.

First, in sensory physiology, research on tuning curves has been extended to naturalistic stimuli and is divided into two complementary approaches: encoding and decoding. Based on models of the stimulus-response function, these approaches rely on the idea that neural activity encodes some features of the external world. Successful reconstructions of complex stimuli based on neural responses (decoding), or successful predictions of responses to new stimuli (encoding) are viewed as proofs that the neural code has been cracked. Interpreting these results in the light of Brette's arguments seems necessary. Initially, the stimulus reconstruction method (decoding) was performed either with simple artificial stimuli (Bialek et al. 1991) or in peripheral sensory systems (Rieke et al. 1995; Warland et al. 1997). More recently, studies have reconstructed natural stimuli from cortical responses (Akbari et al. 2019; Miyawaki et al. 2008; Naselaris et al. 2009), opening the spectacular expectation to read subjects' percepts. In the auditory modality, encoding models were used to investigate neural selectivity to a variety of acoustic properties such as phonetic features (Mesgarani et al. 2014), pitch (Oxenham 2018), and timbre and rhythm (Woolley et al. 2009). To achieve good performance, the stimulus/response models used in decoding/encoding approaches rely on features such as trial averaging, statistics of natural stimuli, and starting time of the stimulus. Thus, the right interpretation should be that an "ideal observer" with *a priori* knowledge of the experimental design can infer the stimulus (in the decoding approach) or the neural response (in the encoding approach). Noteworthy, this field has led to an interesting drift from the idea of a fixed relationship between stimulus and neural responses to a more dynamic model, and is now tackling the mechanisms by which sensory responses are modulated by learning, context, and history (Fritz et al. 2005; Holdgraf et al. 2016; Williamson et al. 2016).

Second, the field of neuroprosthetic devices offers demonstrations of causal links between neural code and brain functions. The most successful of these devices, the cochlear implant (CI), operates with blunt stimulations of auditory nerve terminals. Despite a

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

large current spread in the tympanic ramp, the CI allows implanted subjects to have percepts and recover speech understanding. Even though there are huge differences between the normal cochlea and the CI, the fact that CIs restore hearing can be viewed as a proof that the neural code at play in the periphery has been deciphered and is successfully implemented in a prosthetic device. However, the CI settings that lead to speech comprehension differ considerably from one subject to another, as do the strategies leading to the largest evoked responses in auditory cortex (Adenis et al. 2018). Thus, in contrast to the genetic code that is invariant across cells and species, the neural code (understood as changes in neural activity in adaption to a CI) is probably specific for each individual and/or each type of neuron. In line with sensorimotor theories, the success of CIs shows that the brain is using a new input in a way it can interact again with the environment, which might be the basis of hearing restoration.

A third important field investigates the effect of disrupting a particular feature of neural activity on a cognitive skill. In the visual system, disruption of physiological activity in the primate middle temporal area during presentation of moving stimuli biases the perceptive judgment of a behaving animal (Salzman & Newsome 1994; Salzman et al. 1990), thus making the first link between neural code (understood as a pattern of activity of specific neurons) and behavioral performance. More recently, studies performed in the hippocampus have found that disrupting the replay of spiking patterns occurring across neuronal ensembles during the sharp wave ripples profoundly alters the memory of previously acquired information (Ego-Stengel & Wilson 2010; Girardeau et al. 2009). These data reinforce the notion that neuronal activity patterns do correlate with the acquired information. More importantly, associating a rewarding stimulation of the medial forebrain bundle with a hippocampal place cell activity induced a place preference at the place cell location (de Lavilléon et al. 2015), demonstrating causal links between a particular place cell's firing rate and a specific location memory. In all these examples, the exact neural activity feature (its firing rate or its temporal spike patterns) correlated with the animal's location is unknown, but causal relationships do exist. Yet, causality is not enough to define a neural code.

Clearly, more caution is necessary when discussing the neural code as overstatements made (Ferster & Spruston 1995; Panzeri et al. 2017) tend to generate the illusions that (1) the same code operates in any sensory and motor system, which is obviously not the case; and (2) the brain's cognitive functions consist of manipulating encoded representations of the world, a theory that is controversial. Does this mean that the concept of neural code should be abandoned or should be used to describe studies linking neural activity to brain function? We believe that the neural code definition should be freed from the notion of representation, and we should clarify what we refer to when investigating the neural mechanism of brain functions.

Quantifying the role of neurons for behavior is a mediation question

Ilenna Simone Jones[a] and Konrad Paul Kording[b]

[a]Department of Neuroscience, University of Pennsylvania, Philadelphia, PA 19104 and [b]Departments of Neuroscience and Bioengineering, University of Pennsylvania, Philadelphia, PA 19104.

ilennaj@pennmedicine.upenn.edu kording@upenn.edu
http://kordinglab.com/people/ilenna_jones/index.html http://koerding.com/

doi:10.1017/S0140525X19001444, e233

Abstract

Many systems neuroscientists want to understand neurons in terms of mediation; we want to understand how neurons are involved in the causal chain from stimulus to behavior. Unfortunately, most tools are inappropriate for that while our language takes mediation for granted. Here we discuss the contrast between our conceptual drive toward mediation and the difficulty of obtaining meaningful evidence.

Arguably the most popular question in systems neuroscience is about mediation: We want to know how neurons contribute to the translation from stimuli via groups of neurons to behaviors. Consequently, the field's review papers and discussion sections saliently talk about mediation as do our own papers (Vilares et al. 2012). In systems neuroscience in particular, elucidating the function or role of neurons in circuits, pathways, and networks that mediate behavior is the field's imperative. As Brette points out, neurons are said to represent stimuli with which we also mean that these neurons are important to behavior (Kording et al. 2004; Stevenson et al. 2011; Vilares et al. 2012). Relatedly, neurons are said to encode stimuli with which we mean that they are eventually decoded and hence have a causal impact (Brette, target article; Glaser et al. 2018; Jazayeri & Movshon 2006; Klein et al. 2003). There are only small sections of systems neuroscience that do not aim primarily at causal descriptions. For example, Bayesian psychophysicists often do not make mechanism claims when they point out that behavior is close to optimal (Kording & Wolpert 2004; 2006). Similarly, neuroengineers trying to use brain activity to control a prosthetic device do not need to make causal assumptions (Wolpaw et al. 2000). But by and large, the world of ideas in systems neuroscience is a world of mediation mechanisms and algorithms; it is a world of causality.

When thinking about data, it is natural to think about the events that we are measuring. If we assume a typical recording-only experiment, we have a stimulus *s*, the activity *r* that is recorded, and a behavior *b*, we can make our mediation question clear. Does *r* mediate the influence of *s* on *b*? We can measure the causal effect of *s* on *r*. We can also observe the

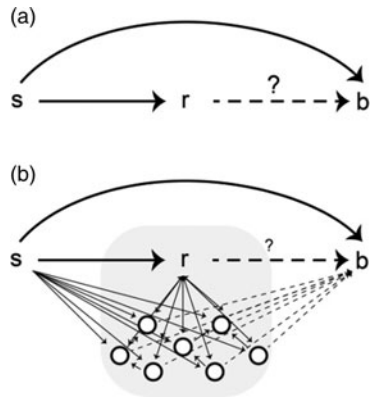


Figure 1. (Jones & Kording) How should we think about mediation? (A) If we record from all the relevant mediators, then we can readily analyze mediation. (B) If we only partially record activities, then an analysis is very complicated.

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

relation between s and b. When only s, r, and b are involved, the causal inference technique called instrumental variable analysis (Angrist & Krueger 2001) allows us to calculate the causal influence of the activity r onto the behavior as

$$CE(r \rightarrow b \mid s) = \frac{Cov(s,\ b)}{Cov(s,\ r)}$$

where the causal effect of r onto b can be determined by the ratio of measured covariances between s and b, and between s and r. For a fixed stimulus, r will have a probability distribution ($p(r|s)$), and the correlation of this with b allows us to estimate the causal mediation effect. The necessary criterion for this reasoning is that there are no causal paths between the variables we reason about that we cannot know. Importantly, if the rest of the brain does not exist, then this way of thinking about mediation analysis is perfectly good. We argue that the way we think about representations and encoding intuitively draws on this idea.

However, we typically record only a tiny proportion of all neurons (Stevenson & Kording 2011). Confounding then makes mediation analysis impossible. Any stimulus-behavior correlation could be due to the neurons we did record or due to other neurons that we did not. Similarly, correlations between neurons can be induced by a paired interaction or indirectly by common input from other neurons. Once our assumptions of full knowledge are violated, our estimates can be arbitrarily off, and our ability to do mediation analysis is gone. We do not learn about the flow of information, or about causal chains, from the kinds of experiments popular in neuroscience.

It is possible to do experiments that get far closer to meaningful claims about mediation. There are four well-established aspects that jointly make mediation more believable.

1. Correlation: Neurons relate to the relevant stimulus aspects and behavior.
2. Necessity: If the neurons are inactivated the behavior is gone.
3. Sufficiency: if the neurons are activated the behavior occurs.
4. Exclusion: The activity is not seen in parallel streams.

Such experiments are beautiful, rare (Kawashima et al. 2016), and generally not doable in typical mammalian settings. However, if we are after mediation effects, then these experiments should be done (Latimer et al. 2015).

Wordings like representation and encoding implicitly suggest that we can arrive at mediation results and, so we argue, are thus popular in neuroscience (Fig 2). Brette points out that this use of language implies that there is a causal relationship between a stimulus and an encoder and between an encoder and an assumed decoder (as in our instrumental variable case). By consistently using words that imply mediation, we are depriving the field of clarity. Language affects the way we formulate models, which in turn affects the experiments we do. As such, it is not just language, but it is the core of what we do as a field.

The focus of the field of mediation analysis may relate to our relative lack of real theories. Mainstream neuroscience theory subscribes to neurons influencing one another and neurons within the same area being similar to one another. But, in a way, those kinds of theories neither do justice to the complex zoo of neural properties nor make the set of possible interpretations of brain data much smaller. If we had meaningful theories, we could test their predictions. Lacking theories, we then simply go for an intuitive mediation analysis, which cannot be well supported by typical experiments. Real theory, including theory that can deal with recurrent systems with circular causality, is needed to break our conceptual reliance of ideas of mediation.

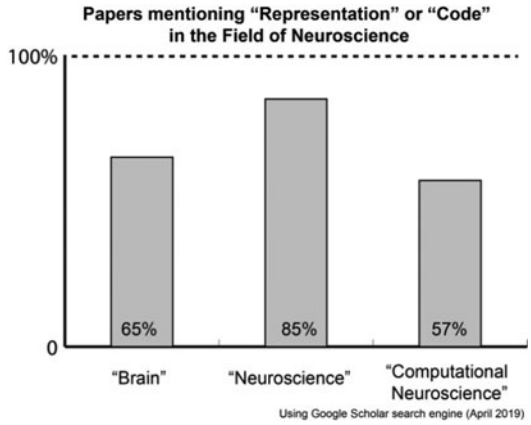


Figure 2. (Jones & Kording) The use of *representation* and *code* is ubiquitous in our field.

Much of what we know about brains comes from the mapping of stimulus-response curves (Wurtz 2009). We were enabled to develop prosthetic devices (Wurtz 2009; Serruya et al. 2002) and new treatments for neurological diseases (Perlmutter & Mink 2006). However, we should not take this impressive story of success as a sign that we do not need to clearly think about what exactly these findings mean. The focus on encoding and representation, if anything, detracts from the importance of the past findings of the field and prevents it from asking how we should think about brains. Moving forward, the field needs to invest in transcending our current theories to make real testable predictions to provide greater precision and logical power to our experiments and understanding of the brain. But tuning curves by themselves can never produce an understanding. After all, we know that theory-free learning about a system is provably impossible (Wolpert & Macready 1997).

Is "the brain" a helpful metaphor for neuroscience?

Fred Keijzer

Department of Theoretical Philosophy, University of Groningen, 9712 GL/19 Groningen,The Netherlands.
f.a.keijzer@rug.nl http://www.rug.nl/staff/f.a.keijzer/index

doi:10.1017/S0140525X19001341, e234

Abstract

Brette criticizes the notion of neural coding as used in neuroscience as a way to clarify the causal structure of the brain. This criticism will be positioned in a wider range of findings and ideas from other branches of neuroscience and biology. While supporting Brette's critique, these findings also suggest the need for more radical changes in neuroscience than Brette envisions.

Brette's analysis and critique of neural coding provide an important challenge for neuroscience. He shows how neuroscience's reliance

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

on neural coding brings in an external scientist's perspective that searches for correlations – encodings – between environmental features that are recognizable by us and neural events made recognizable for us through psychophysical experiments. In this analysis, he argues that finding such correlations between the "outer" world and the "inner" world is of dubious value when it comes to clarifying how nervous systems causally function in organizing behavior. In addition, he stresses the need to maintain the representational sense of the coding metaphor, for example "as a form of normativity realized by anticipatory properties of internal processes" (sect. 5.2, para. 3). In this response, I argue that the issues related to the causal operation of the brain are much more general and pervasive than Brette acknowledges here and that, as a consequence, his commitment to representations becomes problematical.

This wider range of problems becomes visible by highlighting another metaphor, "the brain," a concept widely used – including by Brette – to designate the target of neuroscience. "The brain" may seem a neutral, descriptive term, but just like "neural code," this concept comes with a range of ready-made assumptions, associations, and theoretical commitments that deserve scrutiny.

First, "the brain" refers only to the central, and not the peripheral, nervous system; second, "the" suggests a human brain instead of central nervous systems more generally; and third, "the brain" usually refers to nervous systems as a mental or cognitive control system that consists of an inner information-processing device, linked to the outer world via its sensors and effectors acting as input-output devices. Overall, the phrase "the brain" is associated with interpreting nervous systems in relation to humans, to mind, and to agency.

Using "the brain" to designate the domain of neuroscience is widespread, but it is actually quite strange. It restricts neuroscience to a single (or a few) animal species and comes with a specific and limited focus on what nervous systems might do. It is a bit like developing an account of computation based on a particular computer game. Instead of a human- or mind-oriented view on neuroscience, we require a neuroscience view on neuroscience. I discuss several issues where the difference plays a role.

Large parts of (cognitive) neuroscience still rely on a clear conceptual separation between neurons and other bodily cells, but this *brain-body dualism* is becoming increasingly problematic. For example, many of the molecular ingredients of neurons go back to unicellular ancestors (Sebé-Pedrós et al. 2017); the evolutionary differentiation of neurons from other cells involved a gradual functional segregation (Arendt 2008); and nonneural cells are still involved in electrical signaling, for example, in development (Levin & Martyniuk 2018). Brains are clearly specialized parts of the animal body rather than something of a completely different order.

It is therefore better to talk about "nervous systems" (plural), a phrase that easily accommodates the millions of different types of nervous system organization in existence and refers to whole systems instead of only the central parts. It also applies to relatively "simple" nervous systems that enable fine-grained studies of how relatively small collections of neurons can accomplish sophisticated behavior (e.g., Bargmann & Marder 2013; Liu et al. 2018; Marder 2012).

Focusing on nervous systems more generally provides a different context for neuroscience, which is more concrete than the abstract tasks devised to address behavioral and cognitive questions. Nervous systems are intrinsically active systems that integrate and balance behavioral and perceptual, as well as physiological and developmental processes (Jékely et al. 2015). To understand how nervous systems actually work, a more inclusive view of the various functions of nervous systems is required.

A general neuroscience perspective also takes the peripheral nervous system as an integral part of nervous system functioning. This provides a corrective for the inner-outer dichotomy that looks plausible in a mental context as well as for visual and acoustic input. However, the dichotomy dissolves when it comes to touch and its dependence on self-initiated activity, peripheral nerves, proprioceptive feedback, and the biomechanics of the body itself (Chiel & Beer 1997; Turvey & Fonseca 2014; Tytell et al. 2011). The point can also be made conceptually: The inner-outer dichotomy fits the philosophical notion of mind, but nervous systems are not minds and instead perfectly coextensive with the world, just like the bodies they innervate and the interactions with the world they enable. We do not need to interpret nervous systems as if they are material minds.

To reinforce the last point, it helps to look at recent work on the evolution of the first nervous systems (Kristan 2016). We usually take our own macroscopic view of a world full of trees and animals as basic. From an evolutionary perspective, it is not. Accessing the world at a macroscopic level is a complex achievement that only came about when certain unicellular organisms turned into animals with their nervous systems and complex senses like touch and eyes (Keijzer 2015). Intriguingly, in this context the intuitive idea that nervous systems are at heart input-output devices must be questioned as nervous systems may – at least initially – have acted mostly as internal coordination devices by providing multicellular muscle control (Keijzer & Arnellos 2017; Keijzer et al. 2013) rather than responding to external stimuli.

To conclude, Brette's critique of the coding metaphor as used in neuroscience is spot on and he makes this problem visible in a specific and careful way. However, outside the research field discussed by Brette, many other considerations detract from the idea of neural coding. This wider context furthermore suggests that neuroscience needs ideas that take in, but go beyond viewing the brain as a dynamical system and stressing the centrality of perception-action loops. Similarly, hanging on to representational theorizing is not an obvious way to deal with the many issues raised by a neuroscience view on neuroscience.

Extrinsic and intrinsic representations

Sidney R. Lehky[a] and Anne B. Sereno[b,c]

[a]Computational Neurobiology Laboratory, The Salk Institute, La Jolla, CA 92037; [b]Department of Psychological Sciences, Purdue University, West Lafayette, IN 47907 and [c]School of Biomedical Engineering, Purdue University, West Lafayette, IN 47907.
sidney@salk.edu asereno@purdue.edu
https://engineering.purdue.edu/SerenoLab

doi:10.1017/S0140525X19001262, e235

Abstract

We extend the discussion in the target article about distinctions between extrinsic coding (external references to known things, as required by information theory) and the alternative we and the target article both favor, intrinsic coding (internal relationships within sensory and motor signals). Central to our thinking about intrinsic coding is population coding and the concept of high-dimensional neural response spaces.

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

We are in accord with the view of this target article that neural coding based on information theoretic concepts is insufficient to explain perceptual processing, and extend it in this commentary. Three aspects related to neural codes are discussed in the target article: correspondence, representation, and causality. We focus here primarily on the second one.

To quote from the target article, "Second, neural codes carry information only by reference to things with known meaning. In contrast, perceptual systems must build information from relations between sensory signals and actions, forming an internal model" (abstract). In other words, information theoretic neural codes acquire meaning *externally,* which is problematic as the target article points out, while perceptual systems must acquire meaning *internally* through relationships within the neural system, as the article again points out correctly. We have previously made this distinction between extrinsic and intrinsic neural coding (Lehky et al. 2013), based on similar grounds discussed in the target article.

Population coding was only touched upon lightly in the article. We propose that population codes may be central in implementing intrinsic neural coding. A number of studies on high-level visual processing have interpreted monkey neurophysiological data, using multidimensional scaling (MDS) analysis, in terms of neural populations that form high-dimensional neural response spaces (Baldassi et al. 2013; Eifuku et al. 2004; Kayaert et al. 2005; Kiani et al., 2007; Lehky & Sereno 2007; Murata et al. 2000; Op de Beeck et al. 2001; Rolls & Tovée 1995; Sereno & Lehky 2011; 2018; Sereno et al. 2014; Young & Yamane 1992). In such a neural representation space, each neuron in the population forms one axis of the space. The response to a stimulus is a point in the response space. In other words, the stimulus response is a neural response vector, which, for interpretation using MDS analysis, is simply an ordered list of neural firing rates in the population. No claim is made that the brain is necessarily implementing dimensionality reduction algorithms, but rather dimension reduction offers the possibility of visualizing low-dimensional manifolds existing within high-dimensional neural response space (for discussions, see (Lehky & Sereno 2011; Sereno & Lehky 2011)).

In terms of intrinsic coding, an advantage of using high-dimensional representation spaces is that a stimulus response is indicated by nothing more than raw firing rates in the population. There is no measurement of tuning curves in terms of an externally defined physical parameter or any other externally referenced yardstick. Such external referencing is essentially the fallacy of information-theoretic neural coding, as pointed out in the target article and as exemplified by neural codes mentioned in the target article (Jazayeri & Movshon 2006; Pouget et al. 2003; Quian Quiroga & Panzeri 2009) as well as many others. Rather than external referencing, intrinsic coding uses internal referencing based on relations between stimulus representations. Relationships between stimuli are indicated by geometrical relationships between points (stimulus responses) within the high-dimensional space, such as distances in the neural response space (differences in neural response vectors). Organizing stimulus responses in terms of the neural population vectors themselves provides structure to the responses that go beyond summary statistics of a population response distribution (summary statistics such as are required in the "bag of neurons" model mentioned in the target article).

From our perspective, the way to ascribe meanings to neural representations (i.e., deal with the grounding problem) is through coupled internal representations of sensory and motor variables that become organized and consistent with each other through experience in and interaction with the environment (Lehky et al. 2013). This is essentially the same sensorimotor proposal set forth in the target article. Under our view, populations of neurons form high-dimensional representation spaces, and those representation spaces are organized, developed, and transformed as those populations project to other brain structures, whether they be primarily sensory or motor and whether feedforward recurrent processing, and/or lateral interactions occur. Such an account of neural processing follows a dynamical systems approach and does not incorporate Shannon information theory.

There are a few specific points where we are in disagreement with the target article or feel that amendments are possible. First, with respect to correspondence, we do think that contextual variables can also be encoded with an intrinsic neural population approach and have demonstrated the effects of attention modulation at a single cell in a ventral visual cortical area and its effects at a neural population level on the discriminability of shapes and accuracy of object position (Sereno & Lehky 2018; Sereno et al. 2019).

Second, with respect to causality, we believe the dynamical systems account is consistent with the discrete and timed nature of spikes. Neural processing generally occurs using interactions between analog potentials within the postsynaptic dendritic tree, and not as digital spikes as observed in the axon. Spikes are used for long-distance transmission of the signal along the axon until conversion back to analog form at the next synapse for processing. Dendritic potentials, which are continuous in terms of voltage and time, are the proximate cause of neural processing, not spikes. This view of neural processing as an analog dynamical system is valid regardless of whether spikes can be described by a rate code or time code and regardless of the presence of active potentials within the dendrites.

In sum, we agree with the target article that some currently popular neural coding metaphors are inappropriate and misleading. We agree that these neural coding metaphors are not neutral. They influence the architecture of our conceptual system and limit our understanding. Better metaphors are possible. We propose an alternative intrinsic and relational neural population metaphor.

Codes are for messages, not for neurons

Bjorn Merker

Fjälkestadsvägen 410-82, SE-29194 Kristianstad, Sweden.
bjornmerker@gmail.com

doi:10.1017/S0140525X19001304, e236

Abstract

My commentary draws on extensive arguments against "coding in the brain" developed by my neuroscience mentor, the late Eugene Sachs, who summarized them as follows: "[T]he energy in the signal is the only code there is for information…. The code is the same for each cell, but each cell's location is different, and location is the only basis for significance" (p. 13).

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

The term *coding* along with *processing* and *computation* forms a trio of concepts whose ubiquitous use in neuroscience is underwritten by a usage so loose that often they can replace one another without appreciably affecting the sense of what is being conveyed, as in "We study how the brain codes (or processes, or computes) emotions." As placeholders for "something unknown or yet to be determined," these terms might seem innocuous, were it not for the fact that they carry with them semantic baggage from their nonneural provenance. This technical or formal sense figures in *some*, but far from all, of their uses in neuroscience discourse, a situation that promotes conceptual and communicative confusion.

Regarding the technical and formal sense of "code" and "coding" specifically, the core of the construct as used in the disciplines from which it was imported is that of rule-governed relations of correspondence between two domains with *arbitrary* correspondence assignments in the sense that alternative assignments would work. This is the only common denominator of the two principal codes to be found in nature: (1) the nucleotide-triplet code by which transfer RNA uses the base sequence of messenger RNA to string amino acids into proteins (other genetic operations being template-based), and (2) the two-level combinatorics of phonological and lexical elements by which the sound strings of human languages code the arbitrary (in this case conventional) mapping between the form and meaning of utterances. The arbitrary/conventional element is conspicuous also in the various artificial codes created in reliance on human language: ciphers in cryptography, Morse code in telegraphy, and the many coding schemes employed in the design of digital computers and their programs.

Bona fide codes represent a vanishingly small portion of the myriad lawful relationships that make up the natural world. Likewise for the nervous system: All neural operations of which we have actual knowledge are lawful ones lacking the arbitrary aspect of the correspondence rules of a coding scheme. There is no dearth of appropriate terms for such noncoding lawful relations: "function," "transfer function," "transduction," "mapping," "transform," "representation," and more. Yet the technically incorrect "coding" often substitutes for more informative terms, especially when incompletely known aspects of function, and issues of significance or meaning in particular, are being addressed.

The tacit analogy appears to be the message passing made possible by human language: Neurons "communicate," and somehow the nervous system generates meaning, so perhaps neurons send language-like messages to one another, coding significance or meaning in the temporal sequence of their spike trains. This would supply neurons with an extra-local "code for information" or common language.

But consider the roughly 8000 synapses impinging on a single cortical pyramidal cell: They originate in many hundreds, if not thousands, of other pyramidal cells and subcortical neurons. The blending at the axon initial segment of the graded potentials induced by the synaptic activity of all these sources, excitatory and inhibitory, jointly determines whether that cell will reach threshold to release its all-or-none action potential to its audience of hundreds or thousands of other cells. "Messages" do not survive such treatment anymore than it is possible to monitor the conversations of a cocktail party by a single decibel meter rigged to issue a spike at a fixed sound pressure level. The fact that close to threshold a situation can arise in which the spike output of a neuron replicates the firing of one or a few afferents does not change the basic irrecoverability of the pattern of afferent input from a neuron's output.

Regarding significance, more than half a century of arduous mapping of the response properties of single neurons throughout the nervous system tells us that the principal determinant of what the activity of a neuron signifies is *where it is located*. This "where" ranges from the gross subdivisions of the brain down to a cell's precise connective position in the synaptic network it inhabits (Passingham et al. 2002). Moving a microelectrode to a neighboring cell typically discloses slightly different, but related, response properties, in the aggregate generating the familiar functional maps that abound at every level of the neuraxis.

The reason for this parallel and analog mode of representing significance in the brain is not far to seek. The cortical signal propagation speed of 1 to 6 meters per second (based on Pascual-Leone et al. 2000; Schmolesky et al. 1998) is some 180 million to 30 million times slower than that of electronic circuitry. Operations in this sluggish medium must deliver their global results some three to four times per second (the frequency of gaze movements, the leading edge of most behavior; Merker 2013b). The brain is therefore always strapped for time, and as computer programmers know, when time is short, you don't compute, you use lookup tables.

To perform sophisticated functions with sluggish components in real time, the brain arrays them in complex concatenations of innate as well as acquired (learned) maps ("lookup tables" for "priors") interfaced with one another via a variety of connectivity-based functional transformations. Together they form an analog "computer," not a digital one, for which the analog inner workings of so-called neural networks supply toy models. In analog computers there is no program running on the hardware: The hardware itself *is* the program, and that hardware, moreover, is modifiable by its own activity, generating the learned content of maps.

With no program to run, and no messages to send, there is nothing to code, because significance is not represented in propositional or symbolic form in the brain, but positionally: *where* activity flares is what it signifies or means. But what about the signal discretization of the action potential? A neuron's work takes place by analog blending of graded potentials on its soma-dendrite membrane, and spikes serve only to transmit with fidelity a running record of the upshot of that work to other locations, in keeping with Shannon's insight regarding the utility of discretization for faithful signal *transmission*. No neural work is being performed by discrete spikes except that of bridging distance.

In sum, the answer to the question posed in Romain Brette's title is a resounding "No!"

Encodingism is not just a bad metaphor

Robert Mirski[a,b] and Mark H. Bickhard[b]

[a]The John Paul II Catholic University of Lublin, 20-950 Lublin, Poland and [b]Lehigh University, Bethlehem, PA 18015.
robertmirski@kul.lublin.pl
mhb0@lehigh.edu
https://www.researchgate.net/profile/Robert_Mirski
https://www.lehigh.edu/~mhb0/mhb0.html

doi:10.1017/S0140525X19001286, e237

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

Abstract

Brette's criticism of the coding metaphor focuses on its presence in neurosciences. We argue that this problematic view, which we call "encodingism," is pernicious in *any model of cognition that adopts it*. We discuss some of the more specific problems it begets and then elaborate on Brette's action-based alternative to the coding framework.

Brette argues that encodingism assumptions are pernicious in neuroscience. We would like to expand this critique a bit: Encodingism is a problem in models of cognition in general, not only in neuroscience. We argue that, though encodings certainly exist, they are derivative by nature, and cannot serve to explain the basis of natural cognition. As Brette points out, neurons *could* be said to "encode" the information about some property *for neuroscientists*, as it is they who are interpreting the coding relationship. That is, encodings always require an interpreter who already knows about or represents the two ends of the encoding relationship, as well as the relationship itself. But this representation is exactly the knowledge we are trying to account for when researching minds, and so encodingism becomes circular and leads to an infinite regress of interpretive homunculi. Something else has to lie at the bottom of the natural ability to represent.

The above point underlies Brette's article, but it is also important to note that there is a whole family of problems that plague encodingism. Some of these problems have withstood resolution for millennia. For example, the impossibility of system detectable error that Brette mentions can be traced back to classical scepticism – how am I supposed to know whether what I represent is true, if, in order to find that out, I would have to step outside of myself to gain some independent epistemic access and check? The other end of an encoding is, supposedly, some entity, or property, or state of affairs, but if encodings are all the system has available to represent its reality with, then the only way to attempt to check the encoded end of an encoding is use another encoding. Circularity again.

Foundationalism is another problem forced by encodingism assumptions. Within an encodingist framework, it is impossible for the organism to create first encodings or representations for the very same reason stated above – the organism would have to already know what this particular information is "about" to use it to create a representation. Circularity for a third time.

One would think that this impossibility of representational emergence should automatically discredit encodingism among developmentalists who study the origins of mentality. However, this has not always been the case. Rather, the problem of emergence has been pushed onto biology, and various "core knowledge" accounts have been proposed: infants are supposed to be born with innate theories of physics, biology, or mind (for more criticism, see Allen & Bickhard 2013; Mirski & Gut 2018). But if encodingism blocks emergence in ontogeny, there is no reason why it would not do so in phylogeny too. These are just three of many more problems; for more, see Bickhard and Terveen (1996; Bickhard 2009).

What alternatives are there then? Brette's proposal that we should ask what neurons do rather than what they encode is a significant step in the right direction. However, there are further aspects of cognition, which Brette does not discuss, that we would like to briefly address. Organisms certainly represent reality, and can be wrong about it, and when they are wrong, they often discover that and learn from their mistakes. Naturalism requires that whatever constitutes this representing, and representational error detection, has to emerge at some point from non-representational phenomena. As has been argued, none of these can be accounted for in encodingism in principle, but an action-based perspective has to provide an alternative on pain of being explanatorily vacuous.

Brette briefly mentions what we take to be central to an action-based model when he says "what is useful for the organism is not literally to predict what will happen next, but rather what *might* happen next, conditionally on the actions I can do, so that I can select the appropriate action" (sect. 3.4, para. 6). This statement contains a hint of what mental content can be in an action-based model – the anticipation of possible interactions. This is the proposal of interactivism (Bickhard 2009). (Strictly, it is the anticipation of possible internal process flows that are co-determined by the environment and the organism's actions; it is not anticipation of interaction with the environment as such – there is no surview of the organism and its environment. See, for example, Bickhard 2009; 2015a; 2015b). Such anticipations will have truth value – they implicitly predicate something about the environment (i.e., it is the kind of environment that supports this kind of interaction). And they will be in principle falsifiable and detectable by the organism – all it takes to see if I am right is to actually (try to) engage in the interaction.

As for learning and initial emergence of such action-based normativity, it can happen if we adopt a variation and selection model of learning. If successful anticipations are retained and unsuccessful anticipations are selected against, then the limiting case of representational emergence will be to randomly engage in various interactions and retain the ones that turn out to be successful. No prescience is necessary like in encodingism. Similarly in learning, if my anticipations are falsified, I vary the way I do things until I stumble on a successful alternative. A more detailed discussion of these points can be found elsewhere (Bickhard 2001; 2003; 2009; 2015c; Bickhard & Campbell 1996).

Conceiving of brain functioning in terms of such an anticipatory organization is a viable alternative to coding (Bickhard, 2015a; 2015b). On this view, the brain establishes modes of functioning that implicitly anticipate the upcoming interaction. The modes of functioning are set up by the modulations of such elements as volume transmitters, astrocytes or silent neurons. Such modulations *are* anticipatory in that they set particular modes up, which could turn out to be inappropriate modes for what process flow actually happens. Adopting this alternative, anticipatory view of the brain could complement and extend Brette's proposal. (The above has similarities to some other contemporary frameworks, especially to predictive processing [Clark 2016] and enactivism [Di Paolo & De Jaegher 2012], and indeed there are considerable overlaps, but also fundamental differences [Bickhard 2015b; 2016a; 2016b]).

Acknowledgments. RM was supported by a grant from the National Science Centre (UMO-2016/23/N/HS1/02887).

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

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

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

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

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

indeed the best approach we know of, one perhaps most fruitfully investigated by building "neurorobots," robots controlled by model neuronal systems. This approach has also been called "synthetic neural modeling" (Reeke et al. 1990).

Is coding a relevant metaphor for building AI?

Adam Santoro[1], Felix Hill[1], David Barrett, David Raposo, Matt Botvinick and Timothy Lillicrap

DeepMind, London N1C4AG, United Kingdom.
adamsantoro@google.com felixhill@google.com
barrettdavid@google.com draposo@google.com botvinick@google.com
countzero@google.com www.deepmind.com

doi:10.1017/S0140525X19001365, e240

Abstract

Brette contends that the neural coding metaphor is an invalid basis for theories of what the brain does. Here, we argue that it is an insufficient guide for building an artificial intelligence that learns to accomplish short- and long-term goals in a complex, changing environment.

The goal of neuroscience is to explain how the brain enables intelligent behaviour, while the goal of agent-based artificial intelligence (AI) is to build agents that behave intelligently. Neuroscience, Brette attests, has suffered from an exaggerated (and technically inaccurate) concern for the codes transmitted by particular parts of the brain. In AI, on the other hand, some of the most notable recent progress has been made not by deeply considering neural coding and its implications, but by focusing on higher-level principles from optimization, learning, and control.

Thanks to deep artificial networks trained via backpropagation, we now have artificial learning systems capable of impressive exhibitions of specific human-like skills, such as object recognition and language translation (e.g., He et al. 2016; Vaswani et al. 2017). In artificial, rather than biological, neural networks, we can more tractably characterize the relationship between a model's neural codes, behaviour, and its external "world." AI researchers have full access to their models' input data distribution, can visualise weights and activations in any part of the network and even make causal interventions on them, and can quickly implement new models informed by any coding hypotheses they may have.

Nevertheless, in-depth analysis of a model's internal representations is of increasingly rare concern for getting these models to work. Consider AlphaGo, which is one of the more compelling recent breakthroughs in AI (Silver et al. 2016). Researchers on this project precisely defined the model's goals, the dynamics underlying the model's interactions with its environment, how the model plans its actions, and how the model learns. Each of these components contributes to the model's success, and yet none of them fundamentally depends on considerations from neural coding.

This is not to say that we cannot usefully apply representational analyses to such agents *post hoc*, regardless of whether the representations satisfy Brette's criteria for neural codes (Barack & Jaegle 2019). Indeed, since the earliest days of connectionism, researchers have been interested in the neural codes that emerge when a clearly specified learning algorithm is applied to a well-understood model trained to execute a particular task. A more recent and important collaboration between AI and neuroscience revealed insight into the conditions under which well-known codes can emerge: Grid cells can increasingly be understood as the product of particular optimization processes (Banino et al. 2018; Cueva and Wei 2018). A key feature of these examples, however, is the central descriptive role given to the learning algorithms, architectures, and optimization objectives; neural coding was incidental, and in many cases the codes were not fundamental, privileged primitives on top of which the models were built (Marblestone et al. 2016).

If the broader aim of agent-based AI (Russell & Norvig 2016) is to produce a system that accomplishes short- and long-term goals in a complex, changing environment, then there may be a more pernicious problem to the neural coding framework than it simply being out of vogue in modern AI. How internal responses arrive from given stimuli – a goal that is implicit in the neural coding metaphor – may be logically insufficient for producing intelligent behaviour. In outlining the reasons why, we recall arguments that *any* system – artificial or biological – needs to exert control over its environment to achieve intelligent behaviour.

First, the observations with which an agent may compute do not exist as a prespecified data set, independently of the agent's actions in the world. Rather, it is precisely the decisions that the agent takes in that world that determine the sensory data from which it learns. Second, "[w]ithout an ongoing participation and perception of the world there is no meaning for an agent" (Brooks 1991a, p. 16). An agent participating in an external world that responds to its decisions learns useful, reliable, and meaningful interactions (Cisek 1999). It is these meaningful interactions that ground the agent's representations and allow them to be used for understanding and reasoning about its world. Therefore, insofar as neural coding is understood as a framework to help understand a system's internal stimulus-response patterns, it is a logically insufficient framework for designing AI because of its failure to engage with the agent-environment causal loop.

Given these considerations, what, then, can we say about neural coding's role in describing the brain? In neuroscience, we ultimately care about understanding how the brain enables intelligent behaviour. It is often argued that such an understanding cannot come from analyzing low-level, mechanistic details such as neural codes, because "[a] description of neural activity and connections is not synonymous with knowing what they are doing to cause behavior" (Krakauer et al. 2017). For this level of understanding, we need high-level computational and algorithmic theories that embrace agent-environment interactions. The history of AI tells us that the most useful principles, and the richest theoretical insights, emerged from studying control, optimization, and learning processes rather than the particularities of representations or codes (Sutton 2019). A focus on inferring such processes using our increasing quantities of neural data, rather than characterizing neural codes for their own sake, may also be the most productive way of making progress on understanding intelligent behaviour in humans and animals.

Note

1. AS and FH contributed equally to this work.

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

Neural codes – Necessary but not sufficient for understanding brain function

Simon R. Schultz and Giuseppe P. Gava

Department of Bioengineering and Centre for Neurotechnology, Imperial College London, London SW7 2AZ, United Kingdom.
s.schultz@imperial.ac.uk giuseppe.gava12@imperial.ac.uk

doi:10.1017/S0140525X1900147X, e241

Abstract

Brains are information processing systems whose operational principles ultimately cannot be understood without resource to information theory. We suggest that understanding how external signals are represented in the brain is a necessary step towards employing further engineering tools (such as control theory) to understand the information processing performed by brain circuits during behaviour.

The central tenet of Brette's article is that the notion of a "neural code" is a metaphor, which, despite being popular, does not provide a valid basis for theories of brain function. While we agree with many things that Brette says – and, in particular, would argue that knowing the neural code is not *sufficient* for an understanding of brain function – we think that this central tenet is itself unfounded. A code is, quite simply, a set of rules. One could interpret the "neural code" as the set of rules which neurons obey (we accept that the term is usually used in a somewhat narrower sense). In cryptography, a code is the set of rules which map interpretable messages into a secret form not interpretable by an enemy agent. The aim of the enemy agent is to recover the interpretable form (decode), without knowing in advance the rules. In sensory physiology, we are often concerned with understanding the rules governing how sensory events in the world map onto neuronal activity. Similarly, we do not know the rules (code) *a priori* – but recovering them, through experiment or otherwise, helps us to understand the system. In one case, the rules are designed by a cryptographer; in the other, they emerged through Darwinian selection. However, this has little bearing upon the task confronting the agent. Far from being a metaphor, we argue that the term *neural code* is literally applicable to the rules governing the relationship between environmental stimuli and neural activity, as well as to other areas of brain function such as the generation of movement. We are thus not employing "code" as a metaphor, except inasmuch as such rules may only approximate far more complex biophysics.

Brette argues that neural codes depend upon experimental context. Again, we disagree. The most relevant experimental context is the full repertoire of natural behaviour; we would perform all experiments in this context if it were feasible. Unfortunately, the world is high dimensional, and this would typically result in inadequate sampling of the variables we wish to study within the lifetime of the experiment (or experimenter). We restrict the context in order to sample sufficiently to make statistically valid conclusions (i.e., use an experimental design), attempting instead to ensure generality by also performing different experiments (e.g., with natural images as well as bars). If the neural code has in fact been completely described, the rulebook used does not depend upon context. We are mystified by the counterview proposed – that perceptual systems "build information from relations between sensory signals and actions" (Brette, abstract). Leaving aside that information cannot be "built" (we assume it is representation that is meant here), this structured internal model would, we expect, correspond exactly to our set of rules, or "code."

We think that, to understand brain function, it is likely to be necessary to understand the neural code, as couched in these terms, perhaps with the use of quantitative techniques from information theory (see, e.g., Panzeri and Schultz 2001; Panzeri et al. 1999). However, we would not argue that it is *sufficient*. As engineers, our view is that the brain is an information processing system, and to understand it, we need to know how the information is represented mechanistically. This merely sets the scene: To "reverse engineer" (disassemble to understand) the system, we also need to understand many other "design principles" (Sterling and Laughlin 2015) – the performance constraints, the organisational principles, and so forth – and finally, the specific computations that occur. To do this, we employ (and develop, where needed) additional branches of engineering, such as control theory. The autonomic nervous system perhaps provides a good illustrative example. Parasympathetic and sympathetic nerve fibres connect the internal organs of the body in a complex control system regulating heart rate, digestion, respiratory rate, and many other internal variables. Control theory would seem to be a fruitful approach to study the operations of this system. A first step towards this would be writing down the control variables, for instance, how is blood oxygenation encoded in neural activity in the parasympathetic nervous system? Understanding the neural code for the signals in this system is a necessary (but not sufficient) part of the process of understanding how the system works. Brette provides a counterexample concerning the heart, in which one might conclude that cardiac cells rate-code for running speed. However, any reasonable neural coding analysis of this system would examine vascular flow rates and pressures, obtaining the correct conclusion.

Brette suggests that it is unclear if neural codes can represent structure, giving cell assemblies as an example and noting that a labelled graph, not just a "bag of neurons" (sect. 2.5, para. 2), is required to do so. There is evidence that cell assemblies do support such structured relations. The inhibition of specific assemblies is enough to erase long-lasting behaviours, such as cocaine-driven place preference (Trouche et al. 2016). Co-activation of such cell assemblies can encode goal-driven behaviours, relating both "words" (e.g., places, rewards, actions) and "syntax" (e.g., causal and contextual relations between "words"). Brette alludes to temporal structure as providing a possible solution to this problem, and in fact, there is mounting evidence that this is the case. As an example, consider sharp-wave ripples (SWRs), synchronous, brainwide, fast oscillatory events originating in the hippocampus. During SWRs, place cells are re-activated, according to their spatial tuning, in time-compressed sequences (Buzsaki 2015). Spike timing is crucial in these "replay" events, which were shown to encode previously experienced spatial trajectories and memories in rats (Maboudi et al. 2018). We do not yet fully understand the "rulebook," or neural code, underpinning such structure, but we would argue that doing so will be crucial to understanding how memories are stored, recalled, and used in brainwide networks.

In summary, for us the "neural code" is the set of rules governing how the activity of neurons relates to the state of the system in which they are situated. Knowing this rulebook is likely to be an essential step in the path towards understanding brain function.

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

How can we play together? Temporal inconsistencies in neural coding of music

Björn Vickhoff

Sahlgrenska Academy, University of Gothenburg, 413 90 Gothenborg, Sweden.
bjorn.vickhoff@aniv.gu.se

doi:10.1017/S0140525X19001298, e242

Abstract

If sensory organs encode environment, this code must be decoded to perception. The currently dominant theory of perception – predictive coding – assumes a "Bayesian decoder," a probability function, which will present (to whom?) an optimal guess, given previous encodings of the environment – old codes testing new codes. Such a process would delay perception noticeably. This is inconsistent with the perception of music, which for several reasons must be direct.

Predictive coding (PC) was first introduced for visual perception (Rao & Ballard 1999). PC theory has since been developed in a series of articles to become an influential theory of perception. In PC theory, perception is said to involve probabilistic testing of contextually based hypotheses against sensory input. Deviances from predictions result in prediction errors, which are transmitted upstream in the hierarchically organized brain. This is supposed to reduce *free energy*, that is, reduce the amount of information needed for perception.

Recently, a PC model of music perception was introduced (Koelsch et al. 2018). This model presumes that *event-related potentials* (ERPs), assessed as EEG, reflect hypothesis testing. The cause of ERPs has been debated. It has, however, been demonstrated that some ERPs are reactions on deviances from predictions and, furthermore, that these reactions occur at several levels in the brain hierarchy (Wacogne et al. 2011). ERP thus seems to provide biological confirmation of PC theory. Koelsch et al. (2018) focus on the ERPs ERAN (early right anterior negativity), and MMN (mismatch negativity). There is a vast literature on ERPs for music, indicating reactions on deviances in rhythm, melody, harmony, structure, and so forth. The ERP reaction time for music varies from 200 to 600 ms, depending on the complexity of the stimulus. This leaves us with five questions:

1. How can we play together if the perception of music is delayed? A musician trying to fit in would be about half a second late.
2. How can the sound wave even be perceived as music if the processing times for rhythm, melody, chords, and structure differ? Most musical events are predictable to some extent. Musical beat, for example, is entrained. Here, *entrainment* is the body's synchronization to the beat. We act *on* the beat. Rhythm thus is perceived directly. It has been demonstrated that periodical sounds produce bursts of gamma oscillations on sound onsets and, furthermore, that these bursts continue when the stimulus is omitted (Snyder & Large 2005; Tal et al. 2017). If the perception of unpredicted tones or chords would be delayed noticeably, they would lag the rhythm. The music would fall apart.
3. Why do we not hear musical predictions? According to PC, predictions, if correct, are not affected by sensory input. The prediction should be what we perceive. If a sound is omitted from a predictable pattern, an auditory response can be emitted (Bendixen et al. 2012). Predictions can sometimes be heard as the inner sensation called *musical imagery* (Zatorre et al. 1996). But this is clearly different from actual music. If the band stops playing, we do not hear their music, however predictable it might be.
4. Why do we not hear two tones, when the expectation is violated? A first expected tone should be substituted by an accurate tone. We cannot assume that the second tone mutes the first tone, as the second tone does not exist (in the brain) when the first tone is heard.
5. If it is just an error signal that is sent upstream, the only sensory information about a melody tone is that it is not the expected. But, to infer the actual tone, the brain must know how much and in which direction it deviates from a preceding tone. This is sensory information concerning pitch differences. If we get this information, what is the use of hypothesis testing?

As Brette points out, the word *predictive* in predictive coding does not designate the prediction of future events. Classic PC models are not designed to explain how we perceive time-varying stimuli, as they do not account for neural transmission delays (Hogendoorn & Burkitt 2019). …. Such delays would make it impossible to return a tennis serve, because the actual ball would be several meters ahead of the perceived ball. However, as demonstrated by Nijhawan (1994), the visual system compensates for delays by means of extrapolation of the trajectory of the moving target. This mechanism makes us see the ball where it "should" be, and really is. A striking demonstration of such extrapolation is the flash lag effect (FLE), where a continuously moving object is compared with a discrete repetitive flashing of a stationary cube (https://www.youtube.com/watch?v=DUBM-GG0gAk). It appears as if the moving object is ahead of the flashes, although they are synchronized in real time. It is believed that the movement of the object is extrapolated but not the flashing of the cube; that, thus, the difference in position reflects the difference in neural transmission time. On comparison of FLE to music perception, two observations can be made:

1. In FLE, the perception of the visual pulse (the flashes), although regular and repetitive, is delayed. In music, the perception of the acoustic pulse (the beat) is not.
2. In FLE, time is dissociated (into extrapolated and non-extrapolated time scales), but this is not the case in music.

These points indicate that acoustic perception differs fundamentally from visual perception. Other mechanisms are in play. The Hogendoorn-Burkitt model (Hogendoorn & Burkitt 2019) for visual extrapolation would not work for music, because music is not a continuous movement.

If the function of ERP is not to guide perception, then what? It is possible that ERPs simply reflect negative feedback and that the function is learning – the updating of internal models.

Perception and action are reciprocally dependent. Music is a perfect example. Accordingly, the solution of the dissociation problem may be sought along the lines suggested by Brette, that is, as enactive perception.

Acknowledgment. The author thanks Helge Malmgren for valuable comments.

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

Author's Response

Neural coding: The bureaucratic model of the brain

Romain Brette[a]

[a]Institut de la Vision, Sorbonne Universités, UPMC Univ Paris 06, INSERM, CNRS, 75012 Paris, France.
romain.brette@inserm.fr http://romainbrette.fr

doi:10.1017/S0140525X19001997, e243

Abstract

The neural coding metaphor is so ubiquitous that we tend to forget its metaphorical nature. What do we mean when we assert that neurons encode and decode? What kind of causal and representational model of the brain does the metaphor entail? What lies beneath the neural coding metaphor, I argue, is a bureaucratic model of the brain.

R1. Introduction

Neural coding is a popular metaphor in neuroscience, where objective properties of the world are communicated to the brain in the form of spikes. Most commentators have recognized that the neural coding metaphor is often misused, but they diverge on the extent to which these problems are constitutive of that metaphor.

What is wrong with metaphors (sect. R2)? Metaphors can in principle be useful, as they allow reusing concepts from a different domain. But they can also be misleading when applied to very different domains. Perhaps sensory transduction can be framed as a problem of communication. But are perception and cognition really cases of "world-brain communication" (**Gallistel**)? Unfortunately, this question is rarely explicitly formulated and addressed. Instead, the metaphor captures language and thought in disguise, preempting the meaning of words such as *representation* and *information,* in a way that introduces confusion between the organism's and the observer's perspectives (information for whom?). To understand what lies beneath "neural codes," one must then take a pragmatic approach: How does the neural coding metaphor unfold in reasonings about brain and cognition?

The neural coding metaphor promotes a particular way of understanding causality in complex systems (sect. R3), explanations of the type "A causes B" (e.g., the firing of neuron X causes behavior B). This is an inadequate way of understanding even moderately complex systems of coupled components, such as a system of gears or even a parking lot. In systems, explanations are to be articulated at the level of the organization of processes, not single or even pairs of components. What kind of model of organization features agents that pass formally encoded information along a chain of command with no dynamical constraint? Conceptually, what lies beneath the neural coding metaphor is more than the computer model (**Reeke**): it is a bureaucratic model of the brain.

The neural coding metaphor is tightly linked with the concept of representation, as many commentators have noted (sect. R4). Representation is an important concept, but all supporting arguments are articulated at the level of persons, not neurons – they are considered useful, or necessary to explain certain aspects of cognition. Therefore, those arguments do not entail that representations are neural encodings, as the forms of the bureaucratic model. In fact they cannot be encodings, because encodings need a reader, and then we need to explain how the activity of reading produces an experience with representational content. The way out of the infinite regress is to conceive representation pragmatically in terms of processes with certain properties. This is a challenge the neural coding metaphor covers.

R2. The metaphorical nature of neural codes

R2.1. Is it a metaphor, and what is wrong with metaphors?

Is the "neural code" actually a metaphor? **Schultz & Gava** propose that the neural code is simply "the set of rules which neurons obey" (para. 1) while admitting that the intended meaning is usually more specific, as the target article illustrates. When defined in this very broad way, the terms seem indeed unproblematic. But is it plausible that nothing more is implied when neurons are said to encode stimuli? Schultz & Gava answer themselves negatively: Claiming that the spiking cells of the heart encode running speed in their firing rate is objectionable because it is not a "reasonable coding analysis," since it does not identify the appropriate coding variables. However, what this "reasonable coding analysis" might be in the context of the brain is precisely what is at stake and needs to be defined. Schultz & Gava propose that neural codes are "the rules governing how sensory events in the world map onto neuronal activity" (para. 1), but this does not help understanding why the relation between running speed and firing rates of heart cells is not a reasonable code.

This latter quote is not free of preconceptions: It assumes that the relation between the world and the brain is a mapping (the stimulus-response view), rather than a coupling. As pointed out by **Keijzer**, it pictures the organism as an input-output device (stimulus in, behavior out), rather than an autonomous entity. It rules out the alternative possibility of autonomous neural activity influenced by the environment. Any deviation from the determinism of stimulus responses must then be considered as noise. The notion that the world is mapped to neural activity corresponds to a familiar philosophical position, according to which the brain must hold some copy of the world in order to perceive it. Such philosophical positions deserve exposure and discussion, rather than denial (Andersen et al. 2019).

"A neuron encodes a stimulus" may be presented as a literal description of an experimental observation (a contextual correlation), not a metaphor. But the discourse slips into metaphorical territory every time the brain is claimed to "decode," "read," "interpret," or "manipulate" the neural codes. All commentators who defended neural coding in some form also used a narrower, metaphorical sense. **Gallistel** defends the use of information theory by first framing the problem in terms of "world-brain communication" and considers that "the brain performs arithmetic operations on the signals and symbols" (para. 6). **Gauthier, Loula, Pollock, Wilson, & Wong (Gauthier et al.)** claim that "neural codes must implement the empirically validated representations of computational models" (para. 4). In both cases, the neural code is not just a contextual correlation, it is an atom in a mechanistic model of the brain. **De-Wit, Ekroll, Schwarzkopf, & Wagemans (de-Wit et al.)** agree that the neural coding discourse often improperly focuses on what the

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

experimenter can decode from neural activity (the technical sense), and consider that the important question is what "the brain might be able to decode from that activity" (para. 1). But this more cautious use of the coding metaphor is not free of preconceptions. "Decoding" cannot be literal here since a decoder maps signals to the domain of the original message, not to the biological domain. What then is meant exactly by "decoding" once the observer-centric perspective is rejected?

This imprecision is not without risk. **Merker** complains that "code" is often used improperly. Codes are "rule-governed relations of correspondence between two domains with arbitrary correspondence assignments in the sense that alternative assignments would work" (para. 2), giving the example of the nucleotide-triplet code for amino acids. But the terminology is often applied to any kind of observed relation, creating confusion. Confusion is indeed one risk of metaphor. **Frezza and Zoccolotti** point out that metaphor, including the coding metaphor, is often imprecise and multipatterned, which might explain its success. The danger of metaphor in science, especially when their metaphorical nature is denied, is that key presuppositions are hidden behind the narrative: "The pervasive and persuasive effects of the metaphorical narrative hinder the fundamental self-correcting trait of science that aims to provide counterexamples of dominant theories instead of just supporting them" (last para.). This is because the dominant metaphorical narrative preempts the meaning of words, making it challenging to even articulate an alternative viewpoint. Gibson, for example, while developing a relational view of perceptual information as lawful relations between observables (the "invariant structure" in sensory flow), warned that he used the word *information* for a lack of a better term, and not in the sense of the dominant information processing view (Gibson 1979). The issue is rampant in this discussion about neural coding, because the dominant narrative identifies information with Shannon information and representations with encodings. For example, when I develop an alternative view of information as knowledge built by the organism, in analogy with the way scientific knowledge is built, **Schultz & Gava** object that "information cannot be 'built'" (para. 2), presumably because Shannon information can only decrease with processing. They failed to notice that I attempted to provide a more biologically relevant definition of information, for which the data processing inequality is irrelevant. Similarly, a number of commentators objected to my alleged anti-representational stance (**Birch & Smortchkova**; **Gauthier et al.**; **Huetz, Souffi, Adenis, & Edeline** [**Huetz et al.**]), while others regretted my commitment to representations (**Aranyosi**; **Harnad**; **Keijzer**). What this surprising state of affairs reveals is that representations are identified with encodings, which makes a criticism of encodingism either anti-representationalist or incoherent (see sect. R3).

The great danger of metaphor, when it becomes ubiquitous, is that by preempting the language it also freezes the concepts and hinders critical discussion. As **Jones and Kording** observe, "Language affects the way we formulate models, which in turn affects the experiments we do. As such, it is not just language, but it is the core of what do as a field" (para. 5). Therefore the issue with the neural coding metaphor goes much beyond a matter of terminology. Neuroscientists might use the term *code* in an improper way as **Merker** points out, but this is hardly the major problem at stake. To address the scientific impact of the neural coding metaphor, one must take a pragmatic approach to the meaning of "neural codes," focusing on how they are used in reasonings about brain and cognition.

R2.2. The epistemic danger of the coding metaphor

Several commentators have noted that the coding metaphor promotes a confusion between the experimenter's and the organism's perspectives (**Arsiwalla, Bote, & Verschure** [**Arsiwalla et al.**]; **Gomez-Marin**; **Keijzer**). Gomez-Marin insightfully calls this conflict a "clash of Umwelts." When a correlation between an experimental parameter and neural activity measurements is reported as a "neural code," what the term *code* covers is a relation of command between the experimenter and the organism, where the experimenter imposes a known stimulus onto the observed organism. The "neural code" is about the experimenter's Umwelt, not the organism's Umwelt. Spikes might be signals for the observer, indications that a particular stimulus has been presented. But for the organism, spikes are just the activity of its brain, which obviously depends on the environment it is coupled with, but is not commanded by it; spikes are not necessarily a map of the stimulus world. **Arsiwalla et al.** warn us about "the fallacy of extending conditional epistemic descriptors to ontological explanations of brain and behavior" (last para.): this is what is done every time the brain is presumed to "decode" or "interpret" the "neural code," a construction of the observer's Umwelt.

This confusion of Umwelts leads observers to project their own perspectives onto the organism (**Cao & Rathkopf**; **Gomez-Marin**). As Gomez-Marin puts it, "a description of what the neuroscientist can do prescribes what the animal must do" (para. 1). This is illustrated by **Gauthier et al.**: "the search for neural representations begins with an understanding of the task that an organism solves […] The next step is to propose computational models capable of solving this task" (para.1), and finally "neural codes *must* implement the empirically validated representations of computational models" (para. 4, my emphasis). **Gallistel** describes a similar methodology to study how animals use the sun for navigation: "their brain must subtract the current solar azimuth from the desired compass course to obtain the current solar bearing of the source" (para. 7), concluding that these angles must be encoded by the brain.

This methodology follows Marr's (1982b) classical three levels of analysis: the computational level (what does it do?), the algorithmic level (how does it do it?), the physical level (how is it implemented?), to be studied in this order. The key assumption is that these three levels are independent, an assumption inspired from computers, as **Reeke** and **Frezza & Zoccolotti** have noted. But this independence assumption is not a logical necessity. Bell (1999, p. 2013) has argued convincingly that no such independence exists in the brain: "a computer is an intrinsically dualistic entity, with its physical set-up designed not to interfere with its logical set-up, which executes the computation. In empirical investigation, we find that the brain is not a dualistic entity." In theory also, we find that there is an epistemic problem with the postulate of independence. The brain's algorithms (the second level) are assumed to be based on the manipulation of representations independent from the physical level ("computational objectivism," to use the words of Thompson et al. [1992]). But where do those representations come from, if not the physical level? **Mirski & Bickhard** explain the fallacy of encodingism: "encodings always require an interpreter who already knows about or represents the two ends of the encoding relationship, as well as the relationship itself. But this representation is exactly the knowledge we are trying to account for when researching minds, and so encodingism becomes circular, and leads to an infinite regress of interpretive homunculi" (para. 1).

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

Some commentators partially recognize the epistemic problem of encodings, but remain entrenched in the coding metaphor. **De-Wit et al.** agree that the neural coding discourse often improperly focuses on what the experimenter can decode from neural activity, and conclude that what matters is what "the brain might be able to decode from that activity" (para. 1). But while "decode" literally describes what the experimenter does, it is only applied metaphorically to the brain: The brain does not literally transform its own biological activity into stimulus parameters. And how could one read its own biological activity, if "reading" designates exactly that activity? This weaker version of the coding metaphor does not depart from the observer's perspective. **Lehky & Sereno** agree that there is a conceptual problem with extrinsic codes, defined in reference to something external, as for example tuning curves. They propose two solutions: to replace individual neural responses with high-dimensional population responses, named "population coding," and to consider relations between these high-dimensional vectors, named "intrinsic coding." First, however, with respect to the problem of encodingism, there is no qualitative difference between individual and population responses, and a vector is no more structured than a scalar. Second, to call relations between population responses "intrinsic coding" raises again the observer-organism confusion. It is the observer who notices the relation between high-dimensional responses to stimuli: the responses do not represent the relation, they only instantiate it. Whatever "encodes" these relations is left unexplained. Clearly, it is challenging to conceptualize intrinsic relational representations. But if the goal is to provide an alternative to encodingism, then the temptation to frame them in terms of encodings should be resisted (see sect. R4).

R2.3. Can it be a useful metaphor?

Although metaphors are not literally true, they can still be useful, precisely because they transport familiar concepts to an unfamiliar setting. When are they useful, and when are they pernicious? To take advantage of a metaphor without being carried away by it, one must first acknowledge its metaphorical nature and make its assumptions explicit.

Garson explains that the coding metaphor was central in Adrian's work in the early twentieth century, and proposes that the metaphor was in fact necessary for Adrian to ask questions about the relation between sensory patterns and neural activity patterns and to show, for example, that spike rate increases with stimulus intensity: "It is hard to see how one would even formulate such questions without using the coding metaphor" (para. 6). Yet, such questions are formulated without any allusion to codes in virtually all non-biological domains of science, for example, the relation between atmospheric pressure and rain. Garson correctly notes that the coding metaphor led Adrian to propose the doctrine of "rate coding." It is worth noting that a number of decades later, many have concluded that Adrian has indeed been misled (Brette 2015), including regarding the alleged paradigmatic example of rate coding, neural control of muscular contraction (Sober et al. 2018; Tang et al. 2014; Zhurov and Brezina 2006).

Garson also proposes that the coding metaphor allowed Adrian to ask teleological questions. This is an important remark. One cannot explain organisms without addressing normativity – how is it that an organism can behave appropriately or can live at all, how is it that behavior appears purposeful, and so forth. Normativity is rightfully a key aspect of both the computational program and the efficient coding doctrine. The error is to believe that normativity can only be thought of in terms of codes. On the contrary, the kind of normativity conveyed by the coding metaphor is highly problematic, because it is based on an external reference. Alternative accounts of normativity exist, for example, in enactivism (Maturana and Varela 1973) and interactivism (Bickhard 2009).

Notably, **Santoro**, **Hill**, **Barrett**, **Raposo**, **Botvinick**, **& Lillicrap** (**Santoro et al.**) point out that recent progress in artificial intelligence has generally ignored coding considerations : "the richest theoretical insights, emerged from studying control, optimization, and learning processes rather than the particularities of representations or codes" (last para.). Going further, they observe that the circularity of agent and environment makes it unproductive to think in terms of codes because there is no predetermined set of stimuli to be encoded.

For a communication metaphor to be useful, it should be applied to a problem of communication. We may concede that sensory transduction can be framed in this way (**Gallistel** takes the example of color vision): in order for the organism to be sensitive to electromagnetic waves, this physical dimension must be translated to a biological signal such as ionic currents. It then becomes legitimate to ask questions about signal-to-noise ratio and redundancy, which are indeed questions about the correspondence between two different domains, for which information theory is relevant (Laughlin 1981). Beyond sensory receptors, one may frame the relation between the visual field and the activity of the retinal ganglion cells forming the optic nerve as a communication problem, by noting that the optic nerve creates a bottleneck in a directional flow of excitation (but note that there actually is anatomical feedback to the retina [Gastinger et al. 2006], although from a limited number of neurons). This was essentially Barlow's (1961) motivation when he proposed the efficient coding hypothesis, and presumably what **Harnad** has in mind when he finds it "harmless to call the neural activity along sensory input pathways a 'neural code'" (para. 13).

But the use of the metaphor must still be carefully circumscribed. First, as Barlow noted, viewing the retina through the lens of coding excludes other equally relevant ways to see this system (e.g., as participating in the organism's reaction to specific relevant features [Lettvin et al. 1959]). Second, while the communication metaphor appears adequate when applied to the transformation between physical signals of two different kinds, it becomes much more questionable when the alleged transformation is between properties of things in the world (stimulus parameter or object property) and a biological signal. Do properties of things, such as a category of objects ("trees"), exist as such in the world so that they can be communicated to the brain, or are they abstractions constructed by the mind? If the latter is more accurate, then using a communication metaphor is unproductive.

Even when properly circumscribed, the neural coding metaphor is not without difficulties. Efficient coding offers a normative explanation of transduction in terms of the organism's surroundings (the physical layout of sensory signals), not of its Umwelt (what is meaningful for the organism). This is not a totally irrelevant perspective since the Umwelt depends on the surroundings, but it has limitations. **Gallistel** claims that "the brain's way of encoding color captures a large part of the information available from the reflectance profiles of surfaces in the natural world" (para. 4). Leaving aside the issue that this claim is supported by behavioral rather than physiological evidence, and therefore has little to do with whether and how neurons encode color (a fallacy well described by **Rahnev**), it must be noted that species sharing

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

the same surroundings can have different color vision systems, with different dimensionality, which discards explanations based exclusively on the statistics of natural scenes (Thompson et al. 1992).

Another difficulty has to do with the dynamic aspect of transduction. Many sensory neurons adapt; that is, their firing rate decreases when the stimulus is held constant. Normatively, this allows neurons to remain sensitive to changes in the stimulus. It might be tempting to frame this property as a way to increase (Shannon) information transmitted about the stimulus (Wark et al. 2007). But this raises the observer-organism confusion again: Adaptation can only increase the amount of information if one knows that and how the code changes, but this is only known to the external observer, not the organism who sits at the receiving end. To make such a point, one would need to demonstrate that the organism precisely compensates for the adaptive changes in the code. If the brain metaphorically "decodes" the activity of sensory neurons, then it must be explained how the dynamical and plastic process of coding is perfectly matched to the decoding process in the absence of independent access to the sensory signals.

In summary, the neural coding metaphor can be occasionally useful, if handled with care, but only for a narrow subset of neuroscientific questions. Cognition, in particular, is not a case of "world-brain communication" (**Gallistel**).

R3. Causality in biological systems

As **Jones & Kording** point out, a large part of neuroscience is about understanding how the activity of neurons mediate behavior, that is to say, how neurons are involved in the causal mechanisms underlying behavior. **Huetz et al.** assert that there are "causal links between neural code and brain functions" (para. 3), pointing out that electrically stimulating the auditory nerve produces auditory experience, and electrically stimulating the visual cortex biases visual perception. The issue at stake, however, is not whether neural activity has causal powers – again the neural coding metaphor is so pervasive that "neural code" is identified with neural activity – but whether the causal model that the coding metaphor conveys is correct. **Gulli** contends that I exposed a trivial fallacy, the confusion of correlation and causation: obviously, "A and B are correlated" does not mean "A causes B." He then proposes a checklist of additional tests to establish causality ("causal inferences must be made on the basis of aggregated evidence"). However, my criticism is deeper: In many systems, the relation between two components A and B is simply not of the form A causes B, in which case checklists are irrelevant. I will now give two examples to illustrate this point.

R3.1. The parking lot

The parking lot of an office building has 10 spaces, but there are 12 employees. A few employees complain that they often have to spend time in the morning looking for a parking space in the street. The boss is annoyed: only the employees who arrive late have problems parking. He points out that he arrives very early in the morning and never has any problem finding an empty space: they should stop complaining and get up earlier. Indeed, there is a clear correlation between arrival time and probability of finding an empty space. In addition, if a person decides to arrive earlier then she will find an empty space. Therefore, observation and intervention lead us to concur with the boss that it is the employees' arrival times that cause their ability to find a parking space.

This conclusion is correct in a narrow reductionist sense, that is, the "all else being equal" sense that is relevant to the experimenter's Umwelt. But this sense is essentially irrelevant to understanding how the system works. Not only is it irrelevant, but it is also misleading: Normatively, it leads the boss to conclude that the parking lot can be made to work better by making all the employees arrive earlier, but this is obviously wrong. The parking lot is an example of a system of agents that interact indirectly through the environment, by circular coupling. To understand the system, it is not sufficient to study the relation between an agent and some aspect of the environment. One needs to understand the general organization of the system, the nature of interactions and how they participate to the global function of the system. In other words, one needs a systems approach, not a reductionist ("all else being equal") approach.

Jones & Kording claim that to establish causality, correlation should be supplemented with intervention, and comment that such experiments are "beautiful [but] rare" (para. 4). However, aside from the esthetic aspect, interventional studies do not turn an overly reductionist approach into a more adequate systemic approach (Gomez-Marin 2017; Yoshihara & Yoshihara 2018), and neither does collecting additional "pieces of evidence" as **Gulli** proposes. To understand a complex system using a disparate collection of measurements, the correct approach is not to try to establish causal relations between measurements, but to conceive a model of the system that is consistent with the measurements, focusing on the global organization of the system and its functional logic. As Jones & Kording note: "Real theory, including theory that can deal with recurrent systems with circular causality, is needed to break our conceptual reliance of ideas of mediation" (para. 5).

To deny that components of a system should not be studied as isolated pieces is not to deny that components have a role in the system. **Aranyosi** asserts: "If the reafference and the continuous circular causal loop of organism-environment interaction is truly the ultimate unit of analysis, then there is nothing special about the receptors to consider, or about any other part of the nervous system for that matter" (para. 4). In a systems approach, the "ultimate unit of analysis" is the organization of the system, the relations between components. Therefore the components are important, but the emphasis is on the way they interact.

R3.2. Systems of gears

The brain and environment exhibit circular causality but the coding narrative promotes linear causality. **Barack & Jaegle** object that "linear encoding-decoding relationships between each pair of elements are consistent with an overall picture of a circular, coupled causal system" (para. 8). First, the coding metaphor is not normally applied to each pair of elements (one neuron encodes another neuron?) but to a relation between an external feature and an element (or group of elements). Second, the relation between any two elements might well be linear "all else being equal" (by construction), but studying local interactions in total abstraction of the rest is not a proper way of understanding a system. I will give a second example to illustrate this point.

Consider the system of three gears in Figure R1A. This system had a moment of glory on the internet when the public transport for Greater Manchester decided to put it on an ad with the slogan "Making the city work together." It takes a moment of thought to

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

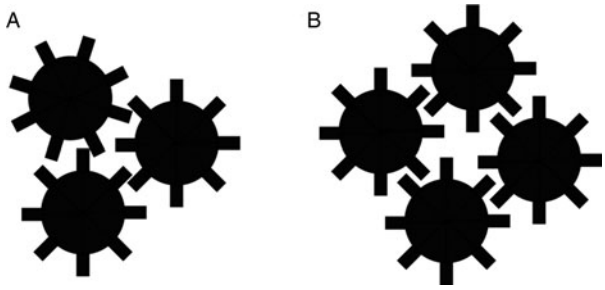


Figure R1. Two systems of gears with different functionality.

realize that gears cannot turn when they are arranged in this way, despite the fact that any two of them fit together and would work in isolation. To understand the difference between a functional (Fig. R1B) and a dysfunctional (Fig. R1A) system of gears, one must go beyond linear interactions between two elements and consider the logic of the system. Mathematically, a functional gear system has a planar bipartite (or two-colorable) graph of contacts (Gordon 1994). B is two-colorable but A is not. If B is a healthy brain and A is a diseased brain, can the coding paradigm help understand why? The argument of approximation offered by **Barack & Jaegle**, that "equivalences exist between dynamical systems with circular causality and approximators with iterated linear causality" (para. 9) misses this point.

R3.3. The bureaucratic model of the brain

What kind of causal model does the coding metaphor promote? In the target article, I argued that the causal structure implied by the coding metaphor is sequential (A causes B, rather than A and B are coupled) and atemporal (timing relations are ignored) and forbids autonomy (B can only result from an external event), three characteristics at odds with the causal structure of biological systems. These are the characteristic features of an algorithm that transforms an input representation into an output representation, by a series of manipulations of intermediate representations. Indeed several commentators have noted the tight relation between the coding metaphor and the computer metaphor (**Reeke**; **Frezza & Zoccolotti**). **Gallistel** uses it extensively to support the coding metaphor: "A computing machine like the brain has […] the machinery for executing operations on symbols" (para. 5). **Gauthier et al.** explicitly consider "brains as *representational* and *computational* devices" (para. 2).

There is a case for the algorithmic model as the underlying causal model of the coding metaphor. This appears in David Marr's influential three levels of analysis of "information processing systems" (Marr 1982b, Fig. 1–4, p. 25). In the "representation and algorithm" level, one should ask "what is the representation for the input and output, and what is the algorithm for the transformation?" Then in the "hardware implementation" level, one should ask "how can the representation and algorithm be realized physically." Marr's view generally fits the computational theory of mind, according to which cognition is the manipulation of symbols by algorithms. The significant leap of faith of the neural coding metaphor is that "neural codes" provide the physical basis ("hardware implementation") of those representations or symbols. But neural codes do not have the quality of symbols: They have a context-dependent meaning and they are abstracted from transient events (spikes), therefore not something that can be manipulated.

However, analyzing the coding metaphor in terms of algorithms makes it difficult to grasp some of the key issues. The fact that different people seem to mean different things about "computer" and "computation" may lead to confusion (Wood 2019). Others might not see what could possibly be wrong with the computer metaphor, since a computer or an algorithm can simulate anything interesting (**Barack & Jaegle**; **Gauthier et al.**). And finally, **Garson** observes that the coding metaphor was used by Adrian in the early twentieth century, well before computers were part of our daily life. Therefore, I suggest that the neural coding metaphor reveals a way to think about causality in complex systems that goes well beyond computer concepts. The coding metaphor sees the brain as a set of agents that communicate information encapsulated in forms along a chain of command. In essence, it is a bureaucratic model of the brain.

A bureaucrat takes an input, and then fills a form. For example it takes an image and fills the form "orientation." Then it passes the form to the next bureaucrat. The bureaucrat will read the forms, apply some rules, and fill some other form, for example, the Jennifer Anniston likeness form. A key feature is that the act of reading has no impact on the form being read (no coupling). Unlike a dynamical system, its activity exists out of time. There are no fixed temporal relations between the different form-filling activities. The bureaucrat outputs a form; the form ends up on the desk of another bureaucrat, who will then process it at some undetermined point. This makes it virtually impossible to explain behavior where a system must interact in real time with its environment. This issue is well described by **Vickhoff** in the context of music perception. Electrophysiological events are often interpreted as encoding sound features, without consideration for the timing of these events. But without time and without temporal coordination, without binding between melody, harmony, and rhythm, there can be no music at all. This is true of all perception but particularly obvious for auditory perception: Percepts are processes that unroll, not forms floating in the brain, waiting to be read.

In the bureaucratic model, the causal structure is essentially sequential, but there can be parallel paths. There can also be feedback: Higher executives can change the forms. **Barack & Jaegle** point out that linear causality between any two elements is not incompatible with circular causality of the overall system. Consider the way the context dependence of neural codes is molded into the coding narrative. Tuning curves in the primary visual cortex (V1) depend on the task being done by the animal (Gilbert & Li 2013); specifically, V1 neurons are sensitive to features important for the task. This effect is described as a "top-down influence," where "top" and "down" refer to the position in the chain of command. The authors correctly note that it raises an issue if we are to think of the activity of V1 neurons as a code for stimulus features, since the meaning of the code would then depend on what the animal is doing. The solution is clear: "The answer lies in the fact that the higher-order areas sent the instruction for these neurons to perform a particular calculation, so the return signal is 'interpreted' by these areas as the result of that calculation and is not confused with other operations those neurons perform." In the bureaucratic model, feedback must be conceptualized as "top-down" instructions for changing the forms. But this bureaucratic concept raises a number of questions: What if the neuron receives feedback from several "higher-order" neurons? Would it not get conflicting instructions? If not, how do the higher-order neurons coordinate themselves? If not by coupling, then who gives instructions to the higher-order neurons?

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

One flaw often attributed to bureaucracies is that they are hopelessly rigid. A bureaucrat has no autonomy: it fills a rigid form instructed from "the top." If the bureaucrat decided to change the form, the result would be disastrous because the rest of the chain applies formal procedures, which would fail. Spontaneous activity is noise, not autonomy. But what should the bureaucrat do when she is supposed to fill the bar orientation form but there is no oriented bar? Or when she is supposed to fill the sound location form but there are two sounds, or the sound of wind? In a real bureaucracy, the stimulus is typically sent back home, or off to some other bureaucrat, but there is no such option for the brain.

Interestingly, while "bureaucracy" tends to evoke an overly rigid and generally dysfunctional mode of organization, there was a time when bureaucracies were seen as efficient ways of organizing work. In the early twentieth century, Max Weber, one of the founders of sociology, was the first to formally study bureaucracies (public or private), and considered that it was the most rational way of organizing work (Weber 1978). All resources are efficiently encoded and processes are designed rationally: What could possibly go wrong?

Coding narratives tend to make extensive use of computational terminology, because the computer metaphor evokes something efficient and powerful. But when we propose that properties are encoded in neural responses, which are then sent to other areas for further processing, the causal model we have in mind is the bureaucratic model of the brain. This model is hard to reconcile with empirical knowledge about the anatomy and physiology of the brain. **Garson** points out that the coding metaphor is used to reason normatively about the brain (what the brain should do to function efficiently). But the situation seems even worse normatively than empirically: Who would think that bureaucracies are a good idealized model of the brain?

R4. Representations

R4.1. Mental representations versus neural encodings

As many commentators have noted, the neural coding metaphor revolves around a central concept in philosophy of mind: representation (**Aranyosi**; **Birch & Smortchkova; Cao & Rathkopf**; **Cisek**; **Deacon & Rączaszek-Leonardi**; **de-Wit et al.**; **Gauthier et al.; Huetz et al.**; **Jones & Kording**; **Keijzer; Lehky & Sereno**; **Mirski & Bickhard**). In fact, only four commentators did not mention it. What are representations and why do many think that they are necessary for cognition (Clark and Toribio 1994)? As Chemero puts it, representations are the "dark matter" of the brain (Chemero 2011, p. 50): they are theoretical constructs considered necessary to explain some features of cognition. One of these features is anticipation: the ability to act as a function of what *might* happen, conditionally on one's actions. In particular, behavior can be directed towards objects that are not present. This is presumably what makes the appeal of predictive coding theory (**Baltieri & Buckley**), despite the fact that it refers to a very narrow notion of anticipation, as I and others have noted (Anderson and Chemero 2013).

More broadly, animals act not only in reaction to proximal stimuli but also as a function of abstract features attributed to sensory signals; these abstract constructions are called internal representations. To take an example from the target article, we could imagine that sound sources can be localized by a simple feedback process: turn the head until the sounds picked up at both ears are equally loud. But that is not what animals generally do, or at least not only. A cat can hear a 100-µs click and then direct its eyes towards the sound source (Populin and Yin 1998), and perceived horizontal sound location is remarkably invariant across large changes in the acoustical signals (Hofman and Van Opstal 1998; Sabin et al. 2005; Yost and Zhong 2014). Even binaural acoustical cues such as interaural time differences vary substantially with the sound's spectrum (Benichoux et al. 2016), but somehow animals behave essentially as a function of an abstract property of the signals, their source's position, and do so while the signals are not present any more. Anti-representationalist views centered on feedback control (Brooks 1991a; Powers 1973a; van Gelder 1998) do not seem to properly address this issue.

This explains why a popular approach to understanding cognition, advocated by **Gauthier et al.**, starts with analyzing how these abstract representations could possibly be extracted from sensory signals (Marr's algorithmic level) and then tries to map this algorithmic process to experimental observables. It is known, for example, that humans and many mammals use mostly intensity differences between the two ears (IIDs) to localize high-frequency sounds in the horizontal plane (Marr's computational level). Therefore, it is thought that, at the algorithmic level, the auditory system computes IIDs and infers sound location from this intermediate calculation. As Gauthier at al. propose, "internal representations [of the computational models] can be used to guide the search for neural codes" (para. 4), and indeed neurons have been identified in the lateral superior olive (LSO) whose firing rate varies monotonically with the IID of an experimental stimulus and therefore "encodes" it. This has formed the consensual view of representation and computation of high-frequency sound localization for several decades: LSO neurons encode IIDs by subtracting the intensity of the two monaural signals. A few authors noted that those neurons are also sensitive to ITDs (Joris and Yin 1995), level and spectrum (Tsai et al. 2010), but the neural coding narrative was compelling. Recently it was found that experimenters had been mistakenly recording interneurons instead of the principal neurons projecting to other areas, which were missed because they fire only transiently to lateralized sounds (Franken et al. 2018). As it turned out, the standard computational model of IID processing was supported by a fiction fueled by the coding narrative, as Bénichoux and Tollin (2018) comment: "The study by Franken et al. is a good example of how prior expectations can involuntarily mislead scientific endeavor."

Gauthier et al. note correctly that the neural coding metaphor guides the search for representations, by helping focus on the "right" candidate representations. But is it a good thing? A critical flaw in the methodology is to implicitly identify mental representations defined at the abstract algorithmic level with neural representations conceptualized as encodings. **Rahnev** clearly explains the fallacy in the context of the Bayesian brain. Arguments supporting the Bayesian brain are based on the allegedly optimal way in which humans behave. Therefore, they support the "as if" view of the theory: People behave as if the brain was performing the computations of Bayesian theory. But when calling the theory Bayesian *brain*, one commits not just to the "as if" view (which is not about the brain) but to the "realist" view, the notion that the brain literally encodes the variables of Bayesian theory and calculates likelihoods. The problem is no argument supports the direct view, only the "as if" view. This realist view is readily endorsed by Gauthier et al.: "neural codes must implement the empirically validated representations of computational models."

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

But "empirically validated" refers to the "as if" view, and therefore, the assertion is not justified. Similarly, **Gallistel** gives the example of animal navigation: "their brain must subtract the current solar azimuth from the desired compass course to obtain the current solar bearing of the source, the angle at which they must hold the sun's image while flying to their destination" (para. 7), but arguments are exclusively based on behavior and therefore no specific conclusion about the brain can be taken. The neural coding metaphor implicitly commits to the "realist" view, which is incoherent, while evidence is provided for the "as if" view, which is not about the brain.

This confusion explains why several commentators have categorized my position as anti-representationalist, despite the fact that one of the main flaws I attributed to neural codes is their lack of representational quality (**Birch & Smortchkova**; **Gauthier et al.**; **Huetz et al.**). Arguments developed in the target article are aimed at the direct view of representations as neural encodings, rather than at mental representations, which are only supported by arguments placed at the level of behavior or cognition. For example, when Clark and Toribio (1994) argue that some problems are "representation-hungry," the argument is based exclusively on behavior and does not rely on any form of encoding. Others regretted my commitment to representations (**Aranyosi**; **Harnad**; **Keijzer**), but this is because representations are identified with encodings, and encodings are (correctly) seen as incoherent or unnecessary (Brooks 1991; Chemero 2011; van Gelder 1995).

Therefore, the debate on representations seems to rely on an implicit identification between mental representations and encodings, promoted by the neural coding metaphor. It can be argued whether "representation" is a good word to designate the fact that cognition and behavior depend on abstract and anticipatory properties of situations. Perhaps it is misleading. The concept, however, is important. Arguably, and although this might sound provocative, Gibson's (1979) affordances are an example of representations in this "as if" sense. In one of my son's child books, a group of different animals stumble on a potty. The frog says: "a bathtub!"; the dog says: "a bowl!"; the mouse says: "a slide!" Animals perceive affordances, anticipatory properties of interaction that depend on their own Umwelt and not just on the physical environment.

Is rejecting encodingism "throwing out the baby with the bathwater" (**Birch & Smortchkova**)? No, because there are ways to conceive these important aspects of representation without neural codes.

R4.2. A short excursion on consciousness

In the target article, I avoided discussing consciousness because it raises many other difficult issues. As **Harnad** correctly points out, strictly speaking, perception refers to conscious experience and it is notoriously hard to explain "how and why organisms feel rather than just do" (last para.). When I used the words perception and percept, I only meant them in the loser sense that is customary in neuroscience, that is, to refer to certain types of tasks (e.g., localizing a sound source).

Nevertheless, our own conscious experience is undoubtedly a chief source of intuition about representations. We believe there are mental representations because at any given moment, it seems that we have access to a sort of subjective "snapshot" of the world, something that is not the physical world but depends on it, in other words a "representation" of the world.

If conscious experience is produced by the brain, then it would seem that there must be a lawful relation between the state of the brain at a given time and the percept that the person is experiencing, in other words an encoding. I will try to show with a simple thought experiment that this intuition is misleading.

In the TV series *Bewitched*, Samantha the housewife twitches her nose and everyone freezes except her. Then she twitches her nose and everyone unfreezes, without noticing that anything happened. For them, time has effectively stopped. Was anyone experiencing anything during that time? According to the encoding view of conscious experience, yes: One experiences the same percept during the entire time, determined by the unchanging state of the brain. But this seems wrong, and indeed in the TV series the characters behave as if there had been no experience at all during that time. The encoding view of conscious experience is wrong because experiencing or perceiving is an activity, not something to be looked at ("by whom?" **Harnad** asks). Therefore, if we are to keep the concept of representation, it has to be conceived not as an encoding but as a process.

R4.3. Beyond representations as encodings

As Bickhard (2009) argues, the belief in encodingism is rooted in substance metaphysics, which describes reality in terms of things of different kinds (e.g., atoms). For example, the neural coding metaphor sees neural activity as a thing that can be read or manipulated. In contrast, process metaphysics describes reality in terms of processes: "processes have their causal powers in virtue of their organization" (Bickhard 2009, p 553). Bickhard points out that historically, science has progressed by shifting from a substance view to a process view of phenomena. For example, fire is no longer considered caused by phlogiston but by the process of combustion.

When the firing rate of a neuron is called a "neural representation," the neuron's activity is assimilated to a thing that can be manipulated and observed, as if it were a sculpture or a painting. But the neuron's activity is not a thing, as the term indicates: it is a process. An action potential is an event, which appears and disappears immediately and has definite effects on the system. Any biologically relevant concept of representation must respect this dynamical nature.

Of course, not *any* dynamical system is a good model of the brain, as pointed out by **Deacon & Rączaszek-Leonardi**. **Garson** observes that the coding metaphor allows teleological reasoning, which a dynamical system might not include. But as **Arsiwalla et al.** point out, teleology figures prominently in at least one major branch of dynamical systems theory: control theory. The Watt governor, chosen by van Gelder (1995) as an example of an elementary cognitive process that is not computational (meant in the conventional sense of manipulating representations in a series of steps), is a feedback control system. Perceptual control theory (Powers 1973a) sees behavior as the "closed-loop control of what the animal senses" (Arsiwalla et al., para. 2).

Control theory is an interesting perspective on behavior, because it respects the dynamical nature of the organism and the circular relation between organism and environment, and also connects with an important physiological concept, homeostasis. However, it is not without difficulties either, for several reasons. First, the physiological concept of homeostasis has some limitations: The organism actively maintains various quantities within certain viable bounds, but it does not necessarily keep them at a fixed value. On the contrary, it adapts them to the

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

dynamic needs of the organism, so that the concept of "allostasis" has been proposed instead (Sterling 2012). An obvious correction is to allow for dynamic rather than static desired states, but this leads to the second issue: Control typically relies on the paradigm of command, with a controller trying to match a desired state expressed by an external agent, which is not modeled. Therefore, it leaves untouched the question of autonomy. Third, control is classically (but perhaps not necessarily) also entrenched in the coding paradigm, with a variable representing the state of the controlled system and a variable representing the desired state. It is tempting to then postulate that a neuron encodes the sensory variable and some other neuron compares it to the target variable (possibly encoded by another neuron) and issues a command accordingly. But again this is an anthropomorphic projection of our own perspective.

Typically, an engineer would design a sensor in charge of doing a measurement. By this, we mean that the sensor produces a quantity (e.g., an electrical voltage) that is in reliable, invariant correspondence with the physical quantity of interest, in other words, an encoding. But biological organisms do not perform measurements in this sense. First, neurons typically produce spike trains, that is, signals that are highly variable when the stimulus is constant or even absent (as retinal ganglion cells in the dark). Therefore neurons map static physical quantities to dynamic processes, so that a neuron's output at a given time cannot be used as a measure. Only some abstract construction such as the "firing rate," which is not manipulated as such by neurons (which react to individual spikes), might be more stable. But in general, this is not the case either because many sensory neurons adapt to stimuli. The relation between physical quantities and sensory neuron activity is not one of measurement but simply of coupling. There is no physiological signal to be maintained constant or close to a desired state, only dynamic processes: Constancy is to be found at the behavioral level, not at the physiological level.

Consider for example a simple feedback loop such as the stretch reflex: A sensory neuron fires action potentials in response to muscle stretch and excites a motoneuron, which then triggers contractions of the muscle. This acts as a negative feedback loop. To understand this, it is not necessary to look for a neural code of stretch in the sensory neuron and to look for a subtraction performed by the motoneuron. It is sufficient to consider the (spiking) dynamical system formed by the neural circuit together with the muscle, and show that it has a stable fixed point, with dynamical properties that are more desirable when the circuit is connected to the muscle.

The free energy principle (discussed by **Baltieri & Buckley**) makes the same problematic commitment to encodings, because free energy is defined as an information-theoretic function of ungrounded abstract variables, not of physiological processes – note that this is different from the physical concept free energy, which applies to equilibrium thermodynamics, not living systems (Martyushev 2018).

It is possible to conceive organism-environment coupling and homeostasis (considered in a broad sense) in terms of processes rather than encodings. One such conceptual framework in theoretical biology is autopoiesis (Maturana and Varela 1973; Varela et al. 1974): a property of an organization of processes that actively maintain the organization despite continuous change of the substance that composes them (e.g., protein turnover). Beyond homeostasis, several commentators have expressed the idea that representations should be conceived not as encodings but in terms of processes (**Cisek**; **Deacon & Rączaszek-Leonardi**; **Mirski & Bickhard**). Specifically, they develop a *pragmatic* concept of representation, oriented on the effects of spikes rather than on their correlation with external features (Deacon & Rączaszek-Leonardi refer to Peirce). Cisek describes pragmatic representations as follows: "As spikes are a means of directing that flow, their activity perforce corresponds to aspects of the world but also to the organism's needs and its policies for meeting those needs. We could call these "pragmatic representations" – activity that doesn't describe the world but instead mediates interaction with it" (para. 4).

At this point, it is important to recall that arguments in favor of the central role of representations in cognition are all about what representations *allow* (the "as if" view). Therefore they chiefly support the pragmatic view of representation. But what is representational about pragmatic representations? **Mirski & Bickhard** focus on the property of anticipation: "the brain establishes modes of functioning that implicitly anticipate the upcoming interaction" (last para.). In the interactivist model of representation (Bickhard 2009), representations are anticipations of potential interactions and their expected impact on the future course of processes of the system. Perhaps it might be more productive to talk about representational processes than representations.

Arguably, this alternative process-based view of representation and cognition opens more questions and conceptual challenges than it solves. These challenges are hidden, not solved, by the neural coding metaphor.

References

[The letters "a" and "r" before author's initials stand for target article and response references, respectively]

Adenis V., Gourévitch B., Mamelle E., Recugnat M., Stahl P., Gnansia D., Nguyen Y. & Edeline J. M. (2018) ECAP growth function to increasing pulse amplitude or pulse duration demonstrates large inter-animal variability that is reflected in auditory cortex of the guinea pig. *PLoS One*. **13**(8):e0201771. [CH]

Adrian E. D. (1928) *The basis of sensation: The action of the sense organs*. W.W. Norton. [JGar]

Adrian E. D. (1932) *The mechanism of nervous action: Electrical studies of the neurone*. University of Pennsylvania Press. [JGar]

Adrian E. D. & Zotterman Y. (1926) The impulses produced by sensory nerve endings. *Part 2. The response of a single end organ*. Journal of Physiology **61**:157–71. [JGar]

Ahissar E. & Assa E. (2016) Perception as a closed-loop convergence process. *eLife* **5**:12830. doi: 10.7554/eLife.12830. [aRB]

Akbari H., Khalighinejad B., Herrero J. L., Mehta A. D. & Mesgarani N. (2019) Towards reconstructing intelligible speech from the human auditory cortex. *Scientific Reports* *29*;**9**(1):874. [CH]

Allen J. W. P. & Bickhard M. H. (2013) Stepping off the pendulum: Why only an action-based approach can transcend the nativist–empiricist debate. *Cognitive Development* **28**(2):96–133. Available at: https://doi.org/10.1016/j.cogdev.2013.01.002. [RM]

Andersen F., Anjum R.L. & Rocca E. (2019) Philosophical bias is the one bias that science cannot avoid. *eLife* **8**:e44929. doi:10.7554/eLife.44929. [rRB]

Anderson A. A. (2019) Assessing statistical results: Magnitude, precision, and model uncertainty. *The American Statistician* **73** (supp1.):118–21. doi:10.1080/00031305.2018.1537889. [RAG]

Anderson J. R. (1989) A rational analysis of human memory. In: *Varieties of memory and consciousness: Essays in honour of Endel Tulving*, ed. H. L. Roediger III & F. I. M. Craik, pp. 195–210. Lawrence Erlbaum Associates. [JGau]

Anderson M. L. & Chemero T. (2013) The problem with brain GUTs: Conflation of different senses of "prediction" threatens metaphysical disaster. *Behavioral and Brain Sciences* **36**(3):204–205. [arRB]

Angrist J. & Krueger A. (2001) Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives* **15**(4):69–85. [ISJ]

Aranyosi I. (2013) *The peripheral mind: Philosophy of mind and the peripheral nervous system*. Oxford University Press. [IA]

Arendt D. (2008) The evolution of cell types in animals: Emerging principles from molecular studies. *Nature Reviews Genetics* **9**(11):868–882. [FK]

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

Ashby W. R. (1965) *Design for a brain: The origin of adaptive behaviour (Vol. 2)*. Chapman and Hall. [PC]

Ashida G. & Carr C. E. (2011) Sound localization: Jeffress and beyond. *Current Opinion in Neurobiology* **21**(5):745–51. [aRB, GF, ISJ]

Baayen R. H., Davidson D. J. & Bates D. M. (2008) Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* **59**(4):390–412. doi:10.1016/j.jml.2007.12.005. [RAG]

Baldassi C., Alemi-Neissi A., Pagan M., Dicarlo J. J., Zecchina R. & Zoccolan D. (2013) Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons. *PLoS Computational Biology* **9**:e1003167. [SRL]

Ball P. (2011) A metaphor too far. *Nature News*. doi: 10.1038/news.2011.115. Available at: http://www.nature.com/news/2011/110223/full/news.2011.115.html. [GF]

Baltieri M. & Buckley C. L. (2017) An active inference implementation of phototaxis. In: CAL 2017: The Fourteenth European Conference on Artificial Life. *Artificial Life Conference Proceedings* 14:36–43. [MB]

Baltieri M. & Buckley C. L. (2019a) Nonmodular architectures of cognitive systems based on active inference. arXiv preprint arXiv:1903.09542. DOI: 10.1109/IJCNN.2019.8852048 [MB]

Baltieri M. & Buckley C. L. (2019b) PID Control as a process of active inference with linear generative models. *Entropy* **21**(3):257. [MB]

Baluška F. & Mancuso S. (2009) Plant neurobiology: From sensory biology, via plant communication, to social plant behavior. *Cognitive Processing* **10**(1):3–7. [SH]

Banino A., Barry C., Uria B., Blundell C., Lillicrap T., Mirowski P., Pritzel A., Chadwick M. J., Degris T., Modayil J., Wayne G., Soyer H., Viola F., Zhang B., Goroshin R., Rabinowitz N., Pascanu R., Beattie C., Petersen S., Sadik A., Gaffney S., King H., Kavukcuoglu K., Hassabis D., Hadsell R., Kumaran D. & Wayne G. (2018) Vector-based navigation using grid-like representations in artificial agents. *Nature* **557**(7705):429–33. [AS]

Barack D. & Jaegle A. (2019) The role of analysis-by-decomposition in neurocognitive modeling. arXiv preprint. [AS]

Bargmann C. I. & Marder E. (2013) From the connectome to brain function. *Nature Methods* **10**(6):483. [FK]

Barlow H. (1961) Possible principles underlying the transformations of sensory messages. In: *Sensory communication*, ed. W. Rosenblith, pp. 217–34. MIT Press. [arRB]

Barlow H. B., Fitzhugh R. & Kuffler S. W. (1957) Change of organization in the receptive fields of the cat's retina during dark adaptation. *Journal of Physiology* **137**:338–54. [aRB]

Bell A. J. (1999) Levels and loops: The future of artificial intelligence and neuroscience. *Philosophical Transactions of the Royal Society B Biological Sciences* **354**:2013–20. [rRB]

Bendixen A., SanMiguel I., & Schröger E. (2012) Early electrophysiological indicators for predictive processing in audition: A review. *International Journal of Psychophysiology* **83**(2):120–31. [BV]

Benichoux V., Fontaine B., Karino S., Joris P. X. & Brette R. (2015) Neural tuning matches frequency-dependent time differences between the ears. *eLife* **4**:06072. [aRB]

Benichoux V., Rébillat M. & Brette R. (2016) On the variation of interaural time differences with frequency. *Journal of the Acoustical Society of America* **139**(4):1810–21. [arRB]

Benichoux V. & Tollin D. J. (2018) These are not the neurons you are looking for. *eLife* **7**:e39244. doi:10.7554/eLife.39244. [rRB]

Bernard C. (1957) *An introduction to the study of experimental medicine*. Dover. [AG-M]

Bialek W., Nemenman I. & Tishby N. (2001) Predictability, complexity, and learning. *Neural Computation* **13**(11):2409–63. [aRB]

Bialek W., Rieke F., de Ruyter van Steveninck R. R. & Warland D. (1991) Reading a neural code. *Science* **252**:1854–57. [CH]

Bickhard M. H. (2001) Error dynamics: The dynamic emergence of error avoidance and error vicariants. *Journal of Experimental and Theoretical Artificial Intelligence* **13**:199–209. [RM]

Bickhard M. H. (2003) Process and emergence: Normative function and representation. In: *Process theories*, ed. J. Seibt, pp. 121–55. Springer Netherlands. [RM]

Bickhard M. H. (2009) The interactivist model. *Synthese* **166**(3):547–91. Available at: https://doi.org/10.1007/s11229-008-9375-x. [arRB, TWD, GF, RM]

Bickhard M. H. (2015a) Toward a model of functional brain processes: I. Central nervous system functional micro-architecture. *Axiomathes* **25**(3):217–38. Available at: https://doi.org/10.1007/s10516-015-9275-x. [RM]

Bickhard M. H. (2015b) Toward a model of functional brain processes: II. Central nervous system functional macro-architecture. *Axiomathes* **25**(4):377–407. Available at: https://doi.org/10.1007/s10516-015-9276-9. [RM]

Bickhard M. H. (2015c) What could cognition be if not computation … or connectionism, or dynamic systems? *Journal of Theoretical and Philosophical Psychology* **35**(1):53–66. Available at: https://doi.org/10.1037/a0038059. [aRB, RM]

Bickhard M. H. (2016a) Inter- and en- activism: Some thoughts and comparisons. *New Ideas in Psychology* **41**:23–32. Available at: https://doi.org/10.1016/j.newideapsych.2015.12.002. [RM]

Bickhard M. H. (2016b) The anticipatory brain: Two approaches. In: V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence*, ed. V. C. Müller, pp. 259–81. Springer International. [RM]

Bickhard M. H. & Campbell R. L. (1996) Topologies of learning and development. *New Ideas in Psychology* **14**(2):111–56. Available at: https://doi.org/10.1016/0732-118X(96)00015-3. [RM]

Bickhard M. H. & Terveen L. (1996) *Foundational issues in artificial intelligence and cognitive science: Impasse and solution (Advances in psychology, Vol. 109)*. Elsevier/North-Holland. [aRB, RM, GNR]

Block N. (2018) If perception is probabilistic, why does it not seem probabilistic? *Philosophical Transactions of the Royal Society B Biological Sciences* **373**:20170341 Available at: http://rstb.royalsocietypublishing.org/lookup/doi/10.1098/rstb.2017.0341. [DR]

Boker S. M. (1997) A measurement of the adaptation of color vision to the spectral environment. *Psychological Science* **8**(2):130–34. [CRG]

Bolz J. & Gilbert C. D. (1986) Generation of end-inhibition in the visual cortex via interlaminar connections. *Nature* **320**:362–65. [aRB]

Bonabeau E., Theraulaz G., Deneubourg J. L., Aron S. & Camazine S. (1997) Self-organization in social insects. *Trends in Ecology & Evolution* **12**(5):188–93. [aRB]

Brette R. (2010) On the interpretation of sensitivity analyses of neural responses. *Journal of the Acoustical Society of America* **128**(5):2965–72. [aRB]

Brette R. (2012) Computing with neural synchrony. *PLoS Computational Biology* **8**(6):e1002561. [aRB]

Brette R. (2015) Philosophy of the spike: Rate-based vs. spike-based theories of the brain. *Frontiers in Systems Neuroscience* **9**:151. [arRB]

Brette R. (2016) Subjective physics. In: *Closed loop neuroscience*, ed. A. El Hady, pp. 146–70. Academic Press. [aRB]

Brooks R. A. (1991a) Intelligence without representation. *Artificial Intelligence* **47**(1–3):139–59. doi:10.1016/0004-3702(91)90053-M. [arRB, PC, AS]

Brooks R. A. (1991b). New approaches to robotics. *Science* **253**(5025):1227–32. [MB]

Bruineberg J., Kiverstein J. & Rietveld E. (2018) The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese* **195**(6):2417–44. [MB]

Buzsáki G. (2010) Neural syntax: Cell assemblies, synapsembles, and readers. *Neuron* **68**:362–85. [aRB]

Buzsáki G. (2015) Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus*, **25**(10):1073–188. [SRS]

Canguilhem G. (2008) *Knowledge of life*. Fordham University Press. [AG-M]

Cao R. (2012) A teleosemantic approach to information in the brain. *Biology & Philosophy* **27**(1):49–71. [RC]

Caton R. (1875) The electric currents of the brain." In: Proceedings of the Forty-Third Annual Meeting of the British Medical Association, August 1875, Edinburgh. *British Medical Journal* **2**:257. doi:10.1136/bmj.2.765.257. [RAG]

Chanauria N., Bharmauria V., Bachatene L., Cattan S., Rouat J. & Molotchnikoff S. (2018) Sound induces change in orientation preference of V1 neurons: Audio-visual cross-influence. Preprint. bioRxiv:269589. [aRB]

Chang M. B., Ullman T., Torralba A., & Tenenbaum J. B. (2017) A compositional object-based approach to learning physical dynamics. Presented at the 5th International Conference on Learning Representations (ICLR 2017), April 24–26, 2017, Toulon, France. Available at: http://hdl.handle.net/1721.1/112749. [JGau]

Chemero A. (2011) *Radical embodied cognitive science*. MIT Press. [rRB]

Chiel H. J. & Beer R. D. (1997) The brain has a body: Adaptive behavior emerges from interactions of nervous system, body and environment. *Trends in Neurosciences* **20**(12):553–57. [FK]

Chomsky N. (1959) A review of B. F. Skinner's *Verbal Behavior*. *Language* **35**:26–58. [aRB]

Churchland M. M., Cunningham J. P., Kaufman M. T., Ryu S. I. & Shenoy K. V. (2010) Cortical preparatory activity: Representation of movement or first cog in a dynamical machine? *Neuron* **68**(3):387–400. [PC]

Churchland M. M. & Shenoy K. V. (2007) Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. *Journal of Neurophysiology* **97**(6):4235–57. [PC]

Cisek P. (1999) Beyond the computer metaphor: Behaviour as interaction. *Journal of Consciousness Studies* **6**(11/12):125–42. [aRB, PC, AS]

Cisek P. (2006) Integrated neural processes for defining potential actions and deciding between them: A computational model. *Journal of Neuroscience* **26**(38):9761–70. [PC]

Cisek P. (2007) Cortical mechanisms of action selection: the affordance competition hypothesis. *Philosophical Transactions of the Royal Society B Biological Sciences* **362**(1485):1585–99. [PC]

Cisek P. (2019) Resynthesizing behavior through phylogenetic refinement. *Attention Perception & Psychophysics*. Available at: https://doi.org/10.3758/s13414-019-01760-1. [PC]

Cisek P. & Kalaska J. F. (2010) Neural mechanisms for interacting with a world full of action choices. *Annual Review of Neuroscience* **33**:269–98. [PC]

Cisek P. & Thura D. (2018) Neural circuits for action selection. In: *Reach-to-grasp behavior: Brain, behavior, and modelling across the life span*, ed. D. Corbetta & M. Santello, Ch. 5. Routledge. [PC]

Clark A. (1997) *Being there: Putting brain, body, and world together again*. MIT Press. [PC]

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

Clark A. (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Science* **36**:181–204. [aRB]
Clark A. (2015) *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press. [MB]
Clark A. (2016) *Surfing uncertainty: Prediction, action, and the embodied mind/Andy Clark*. Oxford University Press. [RM]
Clark A. & Toribio J. (1994) Doing without representing? *Synthese* **101**:401–31. doi:10.1007/BF01063896. [rRB]
Constantinidis C. & Klingberg T. (2016) The neuroscience of working memory capacity and training. *Nature Reviews Neuroscience* **17**:438–49. [aRB]
Cook R., Bird G., Catmur C., Press C. & Heyes C. (2014) Mirror neurons: From origin to function. *Behavioral and Brain Sciences* **37**(2):177–92. [JB]
Crick F. (1979) Thinking about the brain. *Scientific American* **241**:219–32. [aRB]
Cueva C. J. & Wei X. X. (2018) Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. arXiv preprint arXiv:1803.07770. [AS]
Cummins R. (1975) Functional analysis. *Journal of Philosophy* **72**(20):741–65. [DB]
Deacon T. W. (2018) Beneath symbols: Convention as a semiotic phenomenon. In: *Evolution & Contextual Behavioral Science: A Reunification*. ed. S. C. Hayes & D. S. Wilson. New Harbinger. [TWD]
deCharms R. C. & Zador A. (2000) Neural representation and the cortical code. *Annual Review of Neuroscience* **23**:613–47. [aRB]
Deco G., Jirsa V. K. & McIntosh A. R. (2011) Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nature Reviews Neuroscience* **12**:43–56. [aRB]
de Lavilléon G., Lacroix M. M., Rondi-Reig L. & Benchenane K. (2015) Explicit memory creation during sleep demonstrates a causal role of place cells in navigation. *Nature Neuroscience* **18**(4):493–95. [CH]
Dennett D. C. (1978) Why not the whole iguana? *Behavioral and Brain Sciences* **1**:103–04. [aRB]
Dennett D. C. (1981) *Brainstorms: Philosophical essays on mind and psychology*: MIT Press. [DB]
Dewey J. (1896) The reflex arc concept in psychology. *Psychological Review* **3**(4), 357–70. [aRB, PC]
de-Wit L., Alexander D., Ekroll V. & Wagemans J. (2016) Is neuroimaging measuring information in the brain? *Psychonomic Bulletin & Review* **23**(5):1415–28. [Ld-W]
Di Paolo E. & De Jaegher H. (2012) The interactive brain hypothesis. *Frontiers in Human Neuroscience* **6**:163. Available at: https://doi.org/10.3389/fnhum.2012.00163. [RM]
diSessa A. & Abelson H. (1981) *Turtle geometry*. MIT Press. [AG-M]
Dodds W. J. (1878) Localisation of functions of the brain: Being an historical and critical analysis of the question. *Journal of Anatomy and Physiology* **12**(Pt. 4):636–60. [RAG]
Dretske F. (1981) *Knowledge and the flow of information*. MIT Press. [JB, GF]
Dretske F. (1994) If you can't make one, you don't know how it works. *Midwest Studies in Philosophy* **19**:468–82. [RC]
Dyer F. C. & Dickinson J. A. (1996) Sun-compass learning in insects: Representation in a simple mind. *Current Directions in Psychological Science* **5**:67–72. [CRG]
Eccles J. C. (1965) Conscious experience and memory. In: *Brain and conscious experience*, pp. 314–44. Springer. Available at: https://link.springer.com/chapter/10.1007/978-3-642-49168-9_14 [Accessed May 22, 2018]. [aRB]
Eckert R. (1972) Bioelectric control of ciliary activity. *Science* **176**:473–81. [aRB]
Eckert R. & Naitoh Y. (1970) Passive electrical properties of *Paramecium* and problems of ciliary coordination. *Journal of General Physiology* **55**:467–83. [aRB]
Ego-Stengel V. & Wilson M. A. (2010) Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat. *Hippocampus* **20**(1):1–10. [CH]
Eifuku S., De Souza W. C., Tamura R., Nishijo H. & Ono T. (2004) Neuronal correlates of face identification in the monkey anterior temporal cortical areas. *Journal of Neurophysiology* **91**:358–71. [SRL]
Eliasmith C. (2003) Moving beyond metaphor understanding. *The mind for what it is*. The Journal of Philosophy **10**:493–520. [GF]
Eliasmith C. & Anderson C. H. (2004) *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. MIT Press. [JGau]
Erlhagen W. & Schoner G. (2002) Dynamic field theory of movement preparation. *Psychological Review* **109**(3):545–72. [PC]
Fabbri-Destro M. & Rizzolatti G. (2008) Mirror neurons and mirror systems in monkeys and humans. *Physiology* **23**(3):171–79. [JB]
Ferrier D. (1886) *Functions of the brain*, 2nd edition. G. P. Putnam's Sons. [RAG]
Ferster D., & Spruston N. (1995) Cracking the neuronal code. *Science* **270**(5237):756–57. [CH]
Finlay B. L. & Uchiyama R. (2015) Developmental mechanisms channeling cortical evolution. *Trends in Neuroscience* **38**(2):69–76. [PC]
Fiorillo C. D., Kim J. K. & Hong S. Z. (2014) The meaning of spikes from the neuron's point of view: Predictive homeostasis generates the appearance of randomness. *Frontiers in Computational Neuroscience* **8**:49. [XDA]
Fodor J. A. (1990) *A theory of content and other essays*. MIT Press. [RC]
Fodor J. A. & Pylyshyn Z. W. (1988) Connectionism and cognitive architecture: A critical analysis. In: *Connections and symbols*, ed. S. Pinker & J. Mehler. MIT Press. [GNR]
Fragkiadaki K., Agrawal P., Levine S. & Malik J. (2016) Learning visual predictive models of physics for playing billiards. Presented at the International Conference on Learning Representations, San Juan, Puerto Rico. Available at: https://arxiv.org/pdf/1511.07404.pdf. [JGau]
Frank M. J., Seeberger L. C. & O'Reilly R. C. (2004) By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science* **306**(5703):1940–43. [JGau]
Franken T. P., Joris P. X. & Smith P. H. (2018) Principal cells of the brainstem's interaural sound level detector are temporal differentiators rather than integrators. *eLife* **7**: e33854. doi:10.7554/eLife.33854. [rRB]
Friston K. (2009) The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences* **13**:293–301. [aRB]
Friston K. (2010) The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience* **11**:127–38. [aRB]
Friston K. (2011) What is optimal about motor control? *Neuron* **72**(3):488–98. [MB]
Friston K. (2012) The history of the future of the Bayesian brain. *Neuroimage* **62**:1230–33. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22023743. [Accessed March 23, 2019.] [DR]
Friston K., Thornton C. & Clark A. (2012) Free-energy minimization and the dark-room problem. *Frontiers in Psychology* **3**:130. [MB]
Fritz J, Elhilali M, Shamma S. (2005) Active listening: Task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex. *Hearing Research* **206**(1/2):159–76. [CH]
Funahashi K.-i. & Nakamura Y. (1993). Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks* **6**(6):801–06. [DB]
Gallagher S. (2017) *Enactivist interventions: Rethinking the mind*. Oxford University Press. [IA]
Gallistel C.R. (2017). The coding question. *Trends in Cognitive Science* **21**(7):498–508. [CRG]
Gallistel C. R. (2018) Finding numbers in the brain. *Proceedings of the Royal Society London. B Biological Sciences* **373**(1740):20170119. Available at: https://royalsocietypublishing.org/doi/10.1098/rstb.2017.0119. [CRG]
Gallistel C. R. (in press) Where meanings arise and how: Building on Shannon's foundations. *Mind and Language* (invited submission to a special issue on representation). [CRG]
Gallistel C. R. & King A. P. (2010) *Memory and the computational brain: Why cognitive science will transform neuroscience*. Wiley/Blackwell. [CRG]
Gao P., Trautmann E., Yu B., Santhanam G., Ryu S., Shenoy K. & Ganguli S. (2017) A theory of multineuronal dimensionality, dynamics and measurement. bioRxiv 214262. doi: https://doi.org/10.1101/214262. [JGau]
Garson J. (2015) The birth of information in the brain: Edgar Adrian and the vacuum tube. *Science in Context* **28**:31–52. [JGar]
Garson J. (2019) *What biological functions are and why they matter*. Cambridge University Press. [JGar]
Gastinger M. J., Tian N., Horvath T. & Marshak D. W. (2006) Retinopetal axons in mammals: Emphasis on histamine and serotonin. *Current Eye Research* **31**:655–67. doi:10.1080/02713680600776119. [rRB]
Gazzola V. & Keysers C. (2009) The observation and execution of actions share motor and somatosensory voxels in all tested subjects: single-subject analyses of unsmoothed fMRI data. *Cerebral Cortex* **19**(6):1239–55. [XDA]
Geman S., & Geman D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6**(6):721–41. doi:10.1109/TPAMI.1984.4767596. [DB]
Gibson J. J. (1979) *The ecological approach to visual perception*. Routledge. [arRB, PC]
Gilbert C. D. (1996) Plasticity in visual perception and physiology. *Current Opinion in* Neurobiology **6**(2):269–74. [GNR]
Gilbert C. D. & Li W. (2013) Top-down influences on visual processing. *Nature Reviews Neuroscience* **14**:350–63. [arRB]
Girardeau G., Benchenane K., Wiener S. I., Buzsáki G. & Zugaro M. B. (2009) Selective suppression of hippocampal ripples impairs spatial memory. *Nature Neuroscience* **12** (10):1222–23. [CH]
Glaser J. I., Perich M. G., Ramkumar P., Miller L. E. & Kording K. P. (2018) Population coding of conditional probability distributions in dorsal premotor cortex. *Nature Communications* **9**:1788. [ISJ]
Goense J. B. M. & Logothetis N. K. (2008) Neurophysiology of the BOLD fMRI signal in awake monkeys. *Current Biology* **18**(9):631–40. doi:10.1016/j.cub.2008.03.054. [RAG]
Gomez-Marin A. (2017) Causal circuit explanations of behavior: Are necessity and sufficiency necessary and sufficient? In: *Decoding neural circuit structure and function*, ed. A. Çelik & M. F. Wernet, pp. 283–306. Springer. Available at: https://link.springer.com/chapter/10.1007/978-3-319-57363-2_11. [Accessed June 27, 2018.] [XDA, arRB]
Gomez-Marin A. & Mainen Z. F. (2016) Expanding perspectives on cognition in humans, animals, and machines. *Current Opinion in Neurobiology* **37**:85–91. [aRB]
Goodman D. F., Benichoux V. & Brette R. (2013) Decoding neural responses to temporal cues for sound localization. *eLife* **2**(2):e01312. [aRB]
Goodman D. F. M. & Brette R. (2010) Spike-timing-based computation in sound localization. *PLoS Computational Biology* **6**(11):e1000993. [aRB]
Gordon G. (1994) Workable gears, Archimedian solids and planar bipartite graphs. *The American Mathematical Monthly* **101**:527–34. doi:10.2307/2975318. [rRB]

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

Graves A. (2013) Generating sequences with recurrent neural networks. arXiv:1308.0850 [cs.NE]. [DB]
Graziano M. S. (2016) Ethological action maps: A paradigm shift for the motor cortex. *Trends in Cognitive Sciences* **20**(2):121–32. [PC]
Griffiths T. L., Chater N., Norris D. & Pouget A. (2012) How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin* **138**:415–22 Available at: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0026884. [DR]
Grossberg S. (1973) Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics* **52**:213–57. [PC]
Grossberg S. (1978) A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In: *Progress in theoretical biology, Vol. 5*, ed. R. Rosen & F. M. Snell, pp. 233–374. Academic Press. [PC]
Grothe B., Pecka M. & McAlpine D. (2010) Mechanisms of sound localization in mammals. *Physiological Reviews* **90**(3):983–1012. [aRB]
Gulordava K., Bojanowski P., Grave E., Linzen T. & Baroni M. (2018) Colorless green recurrent networks dream hierarchically. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, Vol. 1*, pp. 1195–1205. Association for Computational Linguistics. Available at: https://www.aclweb.org/anthology/N18-1108. [JGau]
Harnad S. (1990a) The symbol grounding problem. In: *Emergent computation*, ed. S. Forrest, pp. 335–46. North-Holland. [GNR]
Harnad S. (1990b) The symbol grounding problem. *Physica D: Nonlinear Phenomena* **42** (1–3):335–46. [aRB, TWD]
Harnad S. (2006) The symbol grounding problem. In: *Encyclopedia of cognitive science*, ed. L. Nadel. Wiley. [SH]
Harnad S. (2009) The annotation game: On Turing (1950) on computing, machinery, and intelligence. In: *Parsing the Turing test*, ed. R. Epstein, G. Roberts & G. Beber. Springer Netherlands. [SH]
Harper N. S. & McAlpine D. (2004) Optimal neural population coding of an auditory spatial cue. *Nature* **430**:682–86. [aRB]
Hartshorne C. & Weiss P., eds. (1931–1963) *Collected papers of Charles Sanders Peirce, Vols. I–VI*. Belknap Press of Harvard University Press. [TWD]
Haugeland J. (1985) *Artificial intelligence*. MIT Press. [RAG]
He K., Zhang X., Ren S. & Sun J. (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–78. IEEE. [AS]
Heess N., Sriram S., Lemmon J., Merel J., Wayne G., Tassa Y., Erez T., Wang Z., Ali Eslami S. M., Riedmiller M. J. & Silver D. (2017) Emergence of locomotion behaviours in rich environments. arXiv:1707.02286 [cs.AI]. [DB]
Hellsten I. (2003) Focus on metaphors: The case of "Frankenfood" on the Web. *Journal of Computer-Mediated Communication* **8**(4):JCMC841. Available at: https://doi.org/10.1111/j.1083-6101.2003.tb00218.x. [GF]
Hendriks-Jansen H. (1996) *Catching ourselves in the act: Situated activity, interactive emergence, evolution, and human thought*. MIT Press. [PC]
Herreros I., Arsiwalla X. D. & Verschure P. (2016) A forward model at Purkinje cell synapses facilitates cerebellar anticipatory control. *Advances in Neural Information Processing Systems* **29**:3828–36. [XDA]
Hill A. B. (1965) The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine* **58**(5):295–300. [RAG]
Hinde R. A. (1966) *Animal behaviour: A synthesis of ethology and comparative psychology*. McGraw-Hill. [PC]
Hofman P. M. & Van Opstal A. J. (1998) Spectro-temporal factors in two-dimensional human sound localization. *Journal of the Acoustic Society of America* **103**:2634–48. [rRB]
Hogendoorn H., & Burkitt A. N. (2019) Predictive coding with neural transmission delays: a real-time temporal alignment hypothesis. *eNeuro* **6**(2):ENEURO.0412-18.2019. [BV]
Hohwy J. (2013) *The predictive mind*, Oxford University Press. [IA, MB]
Hohwy J. (2016) The self-evidencing brain. *Noûs* **50**(2):259–85. [IA]
Holdgraf C. R., de Heer W., Pasley B., Rieger J., Crone N., Lin J. J., Knight R.T., & Theunissen F. E. (2016) Rapid tuning shifts in human auditory cortex enhance speech intelligibility. *Nature Communications* **7**:13654. [CH]
Hosoya T., Baccus S. A. & Meister M. (2005) Dynamic predictive coding by the retina. *Nature* **436**:71–77. [aRB]
Hubel D. H. & Wiesel T. N. (1968) Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology* **195**:215–43. [aRB]
Hudson D. A. & Manning C. D. (2018) Compositional attention networks for machine reasoning. Presented at the International Conference on Learning Representations. arXiv preprint arXiv:1803.03067. [JGau]
Hurley S. (2001) Perception and action: Alternative views. *Synthese* **129**(1):3–40. [aRB]
Izhikevich E. M. (2006) Polychronization: Computation with spikes. *Neural Computation* **18**(2):245–82. doi: 10.1162/089976606775093882. [GNR]
Jazayeri M., & Afraz A. (2017) Navigating the neural space in search of the neural code. *Neuron* **93**(5):1003–14. doi:10.1016/j.neuron.2017.02.019. [JGau, RAG]
Jazayeri M. & Movshon J. A. (2006) Optimal representation of sensory information by neural populations. *Nature Neuroscience* **9**(5):690–96. [aRB, GF, ISJ, SRL]
Jeffress L. A. (1948) A place theory of sound localisation. *Journal of Comparative and Physiological Psychology* **41**(1):35–39. [aRB]
Jékely G., Keijzer F.A. & Godfrey-Smith P. (2015) An option space for early neural evolution. *Philosophical Transactions of the Royal Society B Biological Sciences* **370**:201550181. doi: 10.1098/rstb.2015.0181. [FK]
Jenkins W. M. & Masterton R. B. (1982) Sound localization: Effects of unilateral lesions in central auditory system. *Journal of Neurophysiology* **47**:987–1016. [aRB]
Jennings H. S. (1906) *Behavior of the lower organisms*. Columbia University Press/ Macmillan. Available at: http://archive.org/details/behavioroflowero00jenn. [Accessed December 20, 2015.] [aRB]
Johnson J., Hariharan B., van der Maaten L., Hoffman J., Fei-Fei L., Lawrence Zitnick C. & Girshick R. (2017) Inferring and executing programs for visual reasoning. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2989–2998. IEEE. [JGau]
Jonas H. (2001) *The phenomenon of life*. Northwestern University Press. [AG-M]
Joris P. X., Smith P. H. & Yin T. C. (1998) Coincidence detection in the auditory system: 50 years after Jeffress. *Neuron* **21**(6):1235–38. [aRB]
Joris P. X. & Yin T. C. (1995) Envelope coding in the lateral superior olive. I. Sensitivity to interaural time differences. *Journal of Neurophysiology* **73**:1043–62. [rRB]
Judson H. (1980) *The eighth day of creation*. Simon & Schuster. [CRG]
Kaas J. H. & Stepniewska I. (2016) Evolution of posterior parietal cortex and parietal-frontal networks for specific actions in primates. *Journal of Comparatice Neurology* **524**(3):595–608. [PC]
Kawashima T., Zwart M. F., Yang C.-T., Mensh B. D. & Ahrens M. B. (2016) The serotonergic system tracks the outcomes of actions to mediate short-term motor learning. *Cell* **167**, 933–46.e20. [ISJ]
Kawato M. (1997) Bidirectional theory approach to consciousness. In: *Cognition, computation, and consciousness*, ed. M. Ito, Y. Miyashita & E. T. Rolls. Oxford University Press. [aRB]
Kayaert G., Biederman I. & Vogels R. (2005) Representation of regular and irregular shapes in macaque inferotemporal cortex. *Cerebral Cortex* **15**:1308–21. [SRL]
Keijzer F.A. (2015) Moving and sensing without input and output: Early nervous systems and the origins of the animal sensorimotor organization. *Biology & Philosophy* **30** (3):311–31. doi: 10.1007/s10539-015-9483-1. [FK]
Keijzer F.A. & Arnellos A. (2017) The animal sensorimotor organization: A challenge for the environmental complexity thesis. *Biology & Philosophy* **32**(3):421–41. (doi:10.1007/s10539-017-9565-3). [FK]
Keijzer F. A., Van Duijn M. & Lyon P. (2013) What nervous systems do: Early evolution, input–output, and the Skin Brain Thesis. *Adaptive Behavior* **21**(2):67–84. doi: 10.1177/1059712312465330. [FK]
Keysers C., Kaas J. H. & Gazzola V. (2010) Somatosensation in social perception. *Nature Reviews Neuroscience* **11**(6):417. [XDA]
Kiani R., Esteky H., Mirpour K. & Tanaka K. (2007) Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology* **97**:4296–309. [SRL]
Kirchhoff M. D. (2015) Experiential fantasies, prediction, and enactive minds. *Journal of Consciousness Studies* **22**(3/4):68–92. [IA]
Klein D. J., König P. & Körding K. P. (2003) Sparse spectrotemporal coding of sounds. *EURASIP Journal on Advances in Signal Processing* **2003**;2003:902061. [ISJ]
Knill D. C. & Pouget A. (2004) The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences* **27**(12):712–719 Available at: http://www.ncbi.nlm.nih.gov/pubmed/15541511 [Accessed July 10, 2014]. [aRB, DR]
Koelsch S., Vuust P. & Friston K. (2018) Predictive processes and the peculiar case of music. *Trends in Cognitive Sciences* **23**(1):63–77. [BV]
Kogo N. & Wagemans J. (2013) The "side" matters: How configurality is reflected in completion (Discussion Paper). *Cognitive Neuroscience* **4**:31–61. [Ld-W]
Kolchinsky A. & Wolpert D. H. (2018) Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus* **8**(6):20180041. [MB]
Körding K. P., Kayser C., Einhäuser W. & König P. (2004) How are complex cell properties adapted to the statistics of natural stimuli? *Journal of Neurophysiology* **91**:206–12. [ISJ]
Körding K. P. & Wolpert D. M. (2004) Bayesian integration in sensorimotor learning. *Nature* **427**:244–47. [ISJ]
Körding K. P. & Wolpert D. M. (2006) Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences* **10**:319–26. [ISJ]
Krakauer J. W., Ghazanfar A. A., Gomez-Marin A., MacIver M. A. & Poeppel D. (2017) Neuroscience needs behavior: Correcting a reductionist bias. *Neuron* **93**(3):480–90. [AS]
Kristan Jr. W. B. (2016) Early evolution of neurons. *Current Biology* **26**(20):R949–R954. [FK]
Kumar A., Rotter S. & Aertsen A. (2010) Spiking activity propagation in neuronal networks: reconciling different perspectives on neural coding. *Nature Reviews Neuroscience* **11**(9):615–27. [aRB]
Lake B. M., Ullman T. D., Tenenbaum J. B. & Gershman S. J. (2017) Building machines that learn and think like people. *Behavioral and Brain Sciences* **40**:e253. [XDA]
Lakoff G. & Johnson M. (1980a) *Metaphors we live by*. University of Chicago Press. [aRB, GF]

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

Lakoff G. & Johnson M. (1980b) The metaphorical structure of the human conceptual system. *Cognitive Science* **4**:195–208. [GF]
Latimer K. W., Yates J. L., Meister M. L. R., Huk A. C. & Pillow J. W. (2015) NEURONAL MODELING. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science* **349**:184–87. [ISJ]
Laudanski J., Zheng Y. & Brette R. (2014) A structural theory of pitch. *eNeuro* **1**(1): 0033-14.2014. doi: https://doi.org/10.1523/ENEURO.0033-14.2014. [aRB]
Laughlin S. (1981) A simple coding procedure enhances a neuron's information capacity. Zeitschrift für Naturforschung *C* **36**:910–12. [rRB]
Lebedev M. A. & Wise S. P. (2001) Tuning for the orientation of spatial attention in dorsal premotor cortex. *European Journal of Neuroscience* **13**(5):1002–1008. [PC]
Lehky S. R. & Sereno A. B. (2007) Comparison of shape encoding in primate dorsal and ventral visual pathways. *Journal of Neurophysiology* **97**:307–19. [SRL]
Lehky S. R. & Sereno A. B. (2011) Population coding of visual space: Modeling. *Frontiers in Computational Neuroscience* **4**:155. doi: 10.3389/fncom.2010.00155. [SRL]
Lehky S. R., Sereno M. E. & Sereno A. B. (2013) Population coding and the labeling problem: extrinsic versus intrinsic representations *Neural Computation* **25**:2235–64. [SRL]
Le Mouel C. & Brette R. (2017) Mobility as the purpose of postural control. *Frontiers in Computational Neuroscience* **11**:Article 67. Available at: https://www.frontiersin.org/articles/10.3389/fncom.2017.00067/full. [Accessed June 21, 2018.] [aRB]
Leopold D. A., Mitchell J. F. & Freiwald W. A. (2017) Evolved mechanisms of high-level visual perception in primates. In: *Evolution of Nervous Systems (2nd edition)*, ed. J. Kaas, Vol. 3, pp. 203–35). Academic Press. [PC]
Lettvin J. Y., Maturana H. R., McCulloch W. S. & Pitts W. H. (1959) What the frog's eye tells the frog's brain. *Proceedings of the Institute of Radio Engineers* **47**:1940–51. doi:10.1109/JRPROC.1959.287207. [rRB]
Levin M. & Martyniuk C. J. (2018) The bioelectric code: An ancient computational medium for dynamic control of growth and form. *Biosystems* **164**:76–93. [FK]
Levine S., Finn C., Darrell T. & Abbeel P. (2016) End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research* **17**(1):1334–73. [DB]
Liu H., Yang W., Wu T., Duan F., Soucy E., Jin X. & Zhang Y. (2018) Cholinergic sensorimotor integration regulates olfactory steering. *Neuron* **97**(2):390–405. [FK]
Lycan W. G. (1981) Form, function, and feel. *The Journal of Philosophy* **78**(1):24–50. [DB]
Ma W. J. (2012) Organizing probabilistic models of perception. *Trends in Cognitive Sciences* **16**:511–18 Available at: http://www.ncbi.nlm.nih.gov/pubmed/22981359. [Accessed March 2, 2013.] [DR]
Maboudi K., Ackermann E., de Jong L. W., Pfeiffer B. E., Foster D., Diba K. & Kemere C. (2018) Uncovering temporal structure in hippocampal output patterns. *eLife* 7:ELife.34467. doi:10.7554/ELife.34467. [SRS]
Macmillan N. A. & Creelman C. D. (2005) *Detection theory: A user's guide (2nd edition)*. Lawrence Erlbaum Associates. [aRB]
Maffei G., Herreros I., Sanchez-Fibla M., Friston K. J. & Verschure P. F. (2017) The perceptual shaping of anticipatory actions. *Proceedings of the Royal Society B: Biological Sciences* **284**(1869):20171780. [XDA]
Maloney L.T. (2003) Surface colour perception and environmental constraints. In: *Colour perception: Mind and the physical world*, ed. R. Masufeld & D. Heyer. New York: Oxford. [CRG]
Marblestone A. H., Wayne G. & Kording K. P. (2016) Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience* **10**:94. [AS]
Marder E. (2012) Neuromodulation of neuronal circuits: Back to the future. *Neuron* **76** (1):1–11. [FK]
Marinescu I. E., Lawlor P. N.and Konrad P., Kording K. P. (2018) Quasi-experimental causality in neuroscience and behavioural research. *Nature Human Behaviour* **2**(12):891–98. doi:10.1038/s41562-018-0466-5. [RAG]
Marr D. (1982a) *Vision*. Henry Holt. [DB]
Marr D. (1982b) *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman. [rRB, Ld-W]
Martyushev L. M. (2018) Living systems do not minimize free energy: Comment on "Answering Schrödinger's question: A free-energy formulation" by Maxwell James Dèsormeau Ramstead et al. *Physics of Life Review* **24**:40–41. doi:10.1016/j.plrev.2017.11.010. [rRB]
Mastrogiuseppe F. & Ostojic S. (2018) Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron* **99**(3):609–23.e29. [JGau]
Maturana H. R. & Varela F. J. (1973) *Autopoiesis and cognition: The realization of the living*. D. Reidel. [arRB]
Mausfeld R. (2003) No psychology in–no psychology out. *Psychologische Rundschau* **54** (3):185–91. [Ld-W]
McAlpine D., Jiang D. & Palmer A. R. (2001) A neural code for low-frequency sound localization in mammals. *Nature Neuroscience* **4**:396–401. [aRB]
McClelland J. L. & Rogers T. T. (2003) The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience* **4**(4):310. [JGau]
McLeod C. & Nerlich B., eds. (2018) *Synthetic biology: How the use of metaphors impacts on science, policy and responsible research*. Life Sciences, Society and Policy. Available at: https://www.biomedcentral.com/collections/sbmi. [GF]
Menzel R., Kirbach A., Hass W.-D., Fischer B., Fuchs J., Koblofsky M., Lehmann K., Reiter L., Meyer H., Nguyen H., Jones S., Norton P. & Greggers U. (2011) A common frame of reference for learned and communicated vectors in honeybee navigation. *Current Biology* **21**:645–50. [CRG]
Merker B. (2013a) Cortical gamma oscillations: The functional key is activation, not cognition. *Neuroscience & Biobehavioral Reviews* **37**(3):401–17. [aRB]
Merker B. (2013b) The efference cascade, consciousness, and its self: Naturalizing the first person pivot of action control. *Frontiers in Psychology* **4**:Article 501. [BM]
Mesgarani N., Cheung C., Johnson K. & Chang E. F. (2014) Phonetic feature encoding in human superior temporal gyrus. *Science* **343**(6174):1006–10. [CH]
Millikan R. G. (1984) *Language, thought, and other biological categories: New foundations for realism*. MIT Press. [JB, RC]
Mirski R. & Gut A. (2018) Action-based versus cognitivist perspectives on socio-cognitive development: Culture, language and social experience within the two paradigms. *Synthese* **28**(2):96. https://doi.org/10.1007/s11229-018-01976-y. [RM]
Miyawaki Y., Uchida H., Yamashita O., Sato M. A., Morito Y., Tanabe H. C., Sadato N. & Kamitani Y. (2008) Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron* **60**(5):915–29. [CH]
Młynarski W. & McDermott J. H. (2018) Learning midlevel auditory codes from natural sound statistics. *Neural Computation* **30**(3):631–69. [JGau]
Montague P. R., Dayan P. & Sejnowski T. J. (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience* **16** (5):1936–47. [JGau]
Moser E. I., Kropff E. & Moser M. B. (2008) Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience* **31**(1):69–89. [aRB]
Muckli L., Naumer M. J. & Singer W. (2009) Bilateral visual field maps in a patient with only one hemisphere. *Proceedings of the National Academy of Sciences USA* **106** (31):13034–39. [aRB]
Murata A., Gallese V., Luppino G., Kaseda M. & Sakata H. (2000) Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area AIP. *Journal of Neurophysiology* **83**:2580–601. [SRL]
Naselaris T., Prenger R. J., Kay K. N., Oliver M. & Gallant J. L. (2009) Bayesian reconstruction of natural images from human brain activity. *Neuron* **63**(6):902–15. [aRB, CH]
Neander K. (1995) *Misrepresenting & malfunctioning. Philosophical Studies* **79**(2):109–41. [RC]
Neander K. (2017) *A mark of the mental: In defense of informational teleosemantics*. MIT Press. [JB]
Newen A., De Bruin L. & Gallagher S., eds. (2018) *The Oxford handbook of 4E cognition*. Oxford University Press. [MB]
Nijhawan R. (1994) Motion extrapolation in catching. *Nature* **370**:256–57. [BV]
Niv Y. (2009) Reinforcement learning in the brain. *Journal of Mathematical Psychology* **53** (3):139–54. [JGau]
Noble D. (2008) *The music of life: Biology beyond genes*. Oxford University Press. [aRB]
Olshausen B. A. & Field D. J. (2004) Sparse coding of sensory inputs. *Current Opinion in Neurobiology* **14**(3):481–87. [aRB]
Op de Beeck H., Wagemans J. & Vogels R. (2001) Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience* **4**:1244–52. [SRL]
Oppenheim A. V. & Schafer R. W. (2013) *Discrete-time signal processing (3rd edition)*. Pearson. [DB]
O'Regan J. K. & Noë A. (2001) A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences* **24**(5):939–73. [aRB, CH]
Orlandi N. (2018) Predictive perceptual systems. *Synthese* **195**(6):2367–86. [IA]
Oxenham A.J. (2018) How we hear: The perception and neural coding of sound. *Annual Review of Psychology* **69**:27–50. [CH]
Pakan J. M., Francioni V. & Rochefort N. L. (2018) Action and learning shape the activity of neuronal circuits in the visual cortex. *Current Opinion in Neurobiology* **52**:88–97. [aRB]
Palmer S. E., Marre O., Berry M. J. & Bialek W. (2015) Predictive information in a sensory population. *Proceedings of the National Academy of Sciences USA* **112**:6908–13. [aRB]
Panzeri S., Harvey C. D., Piasini E., Latham P. E., Fellin T. (2017) Cracking the neural code for sensory perception by combining statistics, intervention, and behavior. *Neuron* **93**(3):491–507. [CH]
Panzeri S. & Schultz S. R. (2001) A unified approach to the study of temporal, correlational, and rate coding. *Neural Computation* **13**(6):1311–1349. [SRS]
Panzeri S., Schultz S. R., Treves A. & Rolls E. T. (1999) Correlations and the encoding of information in the nervous system. *Proceedings of the Royal Society B: Biological Sciences* **266**(1423):1001–12. [SRS]
Papineau D. (2003) Is representation rife? *Ratio* **16**(2):107–23. [RC]
Pascual-Leone A., Walsh V. & Rothwell J. (2000) Transcranial magnetic stimulation in cognitive neuroscience – Virtual lesion, chronometry, and functional connectivity. *Current Opinion in Neurobiology* **10**:232–7. [BM]
Passingham R. E., Stephan K. E. & Kötter R. (2002) The anatomical basis of functional localization in the cortex. *Nature Reviews Neuroscience* **3**:606–16. [BM]
Pastor-Bernier A. & Cisek P. (2011) Neural correlates of biased competition in premotor cortex. *Journal of Neuroscience* **31**(19):7083–8. [PC]

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

Pattee H. H. (1973) The physical basis and origin of hierarchical control. In: *Hierarchy theory. The challenge of complex systems*, ed. H. H. Pattee & G. Braziller, pp. 73–108. [TWD]

Pattee H. H. (1997) The physics of symbols and the evolution of semiotic controls. In: *Control mechanisms for complex systems: Issues of measurement and semiotic analysis*, ed. M. Coombs & M. Sulcoski, University of New Mexico Press, pp. 9–25. [TWD]

Pearl J., Glymour M. & Jewell N. P. (2016) *Causal inference in statistics*. Wiley. [RAG]

Perkel D. & Bullock T. (1968) *Neural coding: A report based on an NRP work session*, Neuroscience Research Program Bulletin 6. MIT Press. [aRB]

Perlmutter J. S. & Mink J. W. (2006) Deep brain stimulation. *Annual Review of Neuroscience* **29**:229–57. [ISJ]

Pezzulo G. & Cisek P. (2016) Navigating the affordance landscape: Feedback control as a process model of behavior and cognition. *Trends in Cognitive Sciences* **20**(6):414–24. [aRB]

Phillips C.V. & Goodman K. J. (2004) The missed lessons of Sir Austin Bradford Hill. *Epidemiologic Perspectives & Innovations* **1**:Article 3. doi:10.1186/1742-5573-1-3. [RAG]

Piaget J. (1963) *The origins of intelligence in children*. Norton. [PC]

Polanyi M. (1968) Life's irreducible structure. *Science* **160**:1308–12. doi:10.1126/science.160.3834.1308. [TWD]

Populin L. C. & Yin T. C. T. (1998) Behavioral studies of sound localization in the cat. *The Journal of Neuroscience* **18**:2147–60. [rRB]

Pouget A., Dayan P. & Zemel R. S. (2003) Inference and computation with population codes. *Annual Review of Neuroscience* **26**:381–410. [aRB, GF, SRL]

Powers W. T. (1973a) *Behavior: The control of perception*. Aldine. [arRB, PC]

Powers W. T. (1973b) Feedback: Beyond behaviorism. *Science* **179**(4071):351–6. [XDA]

Pruszynski J. A. & Johansson R. S. (2014) Edge-orientation processing in first-order tactile neurons. *Nature Neuroscience* **17**:1404–9. [IA]

Qian N. & Zhang J. (2019) Neuronal firing rate as code length: A hypothesis. *Computational Brain and Behavior*. Available at: https://doi.org/10.1007/s42113-019-00028-z. [CRG]

Quian Quiroga R. & Panzeri S. (2009) Extracting information from neuronal populations: Information theory and decoding approaches. *Nature Reviews Neuroscience* **10**:173–85. [aRB, SRL]

Quian Quiroga R., Reddy L., Kreiman G., Koch C. & Fried I. (2005) Invariant visual representation by single neurons in the human brain. *Nature* **435**:1102–7. [aRB]

Raczaszek-Leonardi J. (2009) Symbols as constraints: The structuring role of dynamics and self-organization in natural language. *Pragmatics and Cognition* **17**:653–76. doi:10.1075/pc.17.3.09ras. [TWD]

Raczaszek-Leonardi J., Nomikou I., Rohlfing K. J. & Deacon T.W. (2018) Language development from an ecological perspective: Ecologically valid ways to abstract symbols. *Ecological Psychology* **30**:39–73. [TWD]

Rahnev D. & Denison R. N. (2018) Suboptimality in perceptual decision making. *Behavioral and Brain Sciences* **41**:E223. [aRB, DR]

Rajalingham R., Issa E. B., Bashivan P., Kar K., Schmidt K. & DiCarlo J. J. (2018) Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience* **38**(33):7255–69. [JGau]

Rao R. P. & Ballard D. H. (1999) Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience* **2**:79–87. [aRB, BV, IA]

Rathkopf C. (2017) Neural information and the problem of objectivity. *Biology & Philosophy* **32**(3):321–36. [RC]

Reeke Jr. G. N. & Edelman G. M. (1988) Real brains and artificial intelligence. *Daedalus* **117**:143–73. [GNR]

Reeke Jr. G. N., Finkel L. H., Sporns O. & Edelman G. M. (1990) Synthetic neural modeling: A multilevel approach to the analysis of brain complexity. In: *Signal and sense: Local and global order in perceptual maps*, ed. G. M. Edelman, W. E. Gall & W. M. Cowan, pp. 607–706. Wiley. [GNR]

Ricci M., Kim J. & Serre T. (2018) Same-different problems strain convolutional neural networks. Preprint. arXiv1802.03390 Cs Q-Bio. Available at: http://arxiv.org/abs/1802.03390. [Accessed May 28, 2018.] [aRB]

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

Rieke F., Bodnar D. A. & Bialek W. (1995) Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society B Biological Sciences* **262**(1365):259–65. [CH]

Rieke F., Warland D., de Ruyter van Stevenick R. & Bialek W. (1997) *Spikes: Exploring the neural code*. MIT Press. [aRB, CRG, JGar]

Rolls E. T. & Tovée M. J. (1995) Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology* **73**:713–26. [SRL]

Rosen R. (1985) *Anticipatory systems: Philosophical, mathematical and methodological foundations*. Pergamon Press. [aRB]

Roth S. & Black M. J. (2005) Fields of experts: A framework for learning image priors. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* 2:860–7. IEEE. [DB]

Rougier J. (2019) p-Values, Bayes factors, and sufficiency. *The American Statistician* **73** (supp1):148–51. doi:10.1080/00031305.2018.1502684. [RAG]

Rumelhart D. E., McClelland J. L. & The PDP Research Group. (1986) *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations*. MIT Press. [GNR]

Russell S. J. & Norvig P. (2016) *Artificial intelligence: A modern approach*. Pearson Education Limited. [AS]

Ryle G. (1949) *The concept of mind*. University of Chicago Press. [DB]

Sabin A. T., Macpherson E. A. & Middlebrooks J. C. (2005) Human sound localization at near-threshold levels. *Hearing Research* **199**:124–34. doi:10.1016/j.heares.2004.08.001. [rRB]

Sachs E. (circa 1971) *A brain for epistemology* [Unpublished work], p. 13.

Salzman C. D., Britten K. H. & Newsome W. T. (1990) Cortical microstimulation influences perceptual judgements of motion direction. *Nature* **346**(6280):174–7. [CH]

Salzman C. D. & Newsome W. T. (1994) Neural mechanisms for forming a perceptual decision. *Science* **264**(5156):231–7. [CH]

Sanborn A. N. & Chater N. (2016) Bayesian brains without probabilities. *Trends in Cognitive Sciences* **20**:883–93. Available at: http://dx.doi.org/10.1016/j.tics.2016.10.003. [DR]

Santoro A., Hill F., Barrett D., Raposo D., Botvinick M. & Lillicrap T. (2019) Is coding a relevant metaphor for building AI? arXiv:1904.10396 [q-bio.NC]. [DB]

Saxena S. & Cunningham J. P. (2019) Towards the neural population doctrine. *Current Opinion in Neurobiology* **55**:103–11. doi:10.1016/j.conb.2019.02.002. [JGau, RAG]

Schäfer A. M. & Zimmermann H. G. (2007) Recurrent neural networks are universal approximators. In: Artificial Neural Networks – ICANN 2006, ed. S. D. Kollias, A. Stafylopatis, W. Duch & E. Oja. *Lecture Notes in Computer Science*, 4131. [DB]

Schmolesky M. T., Wang Y., Hanes D. P., Thompson K. G., Leutgeb S., Schall J. D. & Leventhal A. G. (1998) Signal timing across the macaque visual system. *Journal of Neurophysiology* **79**:3272–8. [BM]

Schnapf J. L., Kraft T. W. & Baylor D. A. (1987) Spectral sensitivity of human cone photoreceptors. *Nature* **325**:439–41. [aRB]

Schultz W., Dayan P. & Montague P. R. (1997) A neural substrate of prediction and reward. *Science* **275**(5306):1593–9. [JGau]

Searle J.R. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* **3**(3):417–57. [GF, SH]

Sebé-Pedrós A., Degnan B. M. & Ruiz-Trillo I. (2017). The origin of Metazoa: A unicellular perspective. *Nature Reviews Genetics* **18**(8):498. [FK]

Sejnowski T. J., Churchland P. S. & Movshon J. A. (2014) Putting big data to good use in neuroscience. *Nature Neuroscience* **17**(11):1440–1. doi:10.1038/nn.3839. [RAG]

Sereno A. B. & Lehky S. R. (2011) Population coding of visual space: Comparison of spatial representations in dorsal and ventral pathways. *Frontiers in Computational Neuroscience* **4**:159. doi:10.3389/fncom.2010.00159. [SRL]

Sereno A. B. & Lehky S. R. (2018) Attention effects on neural population representations for shape and location are stronger in the ventral than dorsal stream. *eNeuro* **5**: e0371-0317.2018. [SRL]

Sereno A. B., Lehky S. R. & Sereno M.E. (2019) Representation of shape, space, and attention in monkey cortex. [Epub ahead of print] *Cortex*. doi:10.1016/j.cortex.2019.06.005. [SRL]

Sereno A. B., Sereno M. E. & Lehky S. R. (2014) Recovering stimulus locations using populations of eye-position modulated neurons in dorsal and ventral visual streams of non-human primates. *Frontiers in Integrative Neuroscience* **8**:28. doi:10.3389/fnint.2014.00028. [SRL]

Seriès P., Latham P. E. & Pouget A. (2004) Tuning curve sharpening for orientation selectivity: Coding efficiency and the impact of correlations. *Nature Neuroscience* **7**:1129–35. [aRB]

Serruya M. D., Hatsopoulos N. G., Paninski L., Fellows M. R. & Donoghue J. P. (2002) Instant neural control of a movement signal. *Nature* **416**:141–2. [ISJ]

Shackleton T. M., Skottun B. C., Arnott R. H. & Palmer A. R. (2003) Interaural time difference discrimination thresholds for single neurons in the inferior colliculus of guinea pigs. *Journal of Neuroscience* **23**(2):716–24. [aRB]

Shannon C. E. (1948) A mathematical theory of communication. *Bell Systems Technical Journal* **27**(3):379–423, 623–56. Available at: https://archive.org/details/bellsystemtechni27amerrich/page/379. [aRB, TWD, CRG, SH]

Shannon C. E. & Weaver W. (1963) *The mathematical theory of communication*. University of Illinois Press. [DB]

Shea N. (2018) *Representation in cognitive science*. Oxford University Press. [JB, RC]

Silver D., Huang A., Maddison C. J., Guez A., Sifre L., Van Den Driessche G., Schrittwieser J., Antonoglou I., Panneershelvam V., Lanctot M., Dieleman S., Grewe D., Nham J., Kalchbrenner N., Sutskever I., Lillicrap T., Leach M., Kavukcuoglu K., Graepel T. & Hassabis D. (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* **529**(7587):484–9. [AS]

Simoncelli E. P. (2003) Vision and the statistics of the visual environment. Current Opinion in Neurobiology **13**:144–9. [aRB]

Singer W. (1999) Neuronal synchrony: A versatile code for the definition of relations? *Neuron* **24**:49–65. [aRB]

Skottun B. C. (1998) Sound localization and neurons. *Nature* **393**(6685):531. [aRB]

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

Skyrms B. (2010). *Signals: Evolution, learning, and information*. Oxford University Press. [RC]

Smolensky P. & Legendre G. (2006) *The harmonic mind: From neural computation to optimality-theoretic grammar: Vol. 1. Cognitive architecture*. MIT Press. [JGau]

Snyder J. S. & Large E. W. (2005) Gamma-band activity reflects the metric structure of rhythmic tone sequences. *Cognitive Brain Research* **24**(1):117–26. [BV]

Sober S. J., Sponberg S., Nemenman I. & Ting L. H. (2018) Millisecond spike timing codes for motor control. *Trends in Neurosciences* **41**:644–8. doi:10.1016/j.tins.2018.08.010. [rRB]

Somjen G. (1972) *Sensory coding in the mammalian nervous system*. Springer. Available at: www.springer.com/us/book/9781468417074. [Accessed March 26, 2018.] [aRB, GF]

Song L., Langfelder P. & Horvath S. (2013) Random generalized linear model: A highly accurate and interpretable ensemble predictor. *BMC Bioinformatics* **14**(1):5. doi:10.1186/1471-2105-14-5. [RAG]

Srivastava N., Mansimov E. & Salakhutdinov R. (2015) Unsupervised learning of video representations using LSTMs. *Proceedings of Machine Learning Research* **37**:843–52. [DB]

Steinberg E. E., Keiflin R., Boivin J. R., Witten I. B., Deisseroth K. & Janak P. H. (2013) A causal link between prediction errors, dopamine neurons and learning. Nature Neuroscience **16**(7):966–73. [JGau]

Stelmach A. & Nerlich B. (2015) Metaphors in search of a target: The curious case of epigenetics. *New Genetics and Society* **34**(2):196–218. [GF]

Sterling P. (2012) Allostasis: A model of predictive regulation. *Physiology & Behavior* **106**:5–15. doi:10.1016/j.physbeh.2011.06.004. [rRB]

Sterling P. & Laughlin S. (2015) *Principles of neural design*. MIT Press. [rRB, SRS]

Stevens C. F. (2018) Conserved features of the primate face code. *Proceedings of the National Academy of Sciences USA* **115**(3):584–8. doi:10.1073/pnas.1716341115. [CRG]

Stevenson I. H., Cherian A., London B. M., Sachs N. A., Lindberg E., Reimer J., Slutzky M. W., Hatsopoulos N. G., Miller L. E. & Kording K. P. (2011) Statistical assessment of the stability of neural movement representations. *Journal of Neurophysiology* **106**(2):764–74. [ISJ]

Stevenson I. H. & Kording K. P. (2011) How advances in neural recording affect data analysis. *Nature Neuroscience* **14**:139–42. [ISJ]

Sutton R. (2019) The bitter lesson. Available at: http://incompleteideas.net/IncIdeas/BitterLesson.html. [AS]

Suvrathan A., Payne H. L., & Raymond J. L. (2016) Timing rules for synaptic plasticity matched to behavioral function. *Neuron* **92**(5):959–67. [XDA]

Syka J. & Straschill M. (1970) Activation of superior colliculus neurons and motor responses after electrical stimulation of the inferior colliculus. *Experimental Neurology* **28**:384–92. [aRB]

Tal I., Large E. W., Rabinovitch E., Wei Y., Schroeder C. E., Poeppel D. & Golumbic E. Z. (2017) Neural entrainment to the beat: The "missing-pulse" phenomenon. *Journal of Neuroscience* **37**(26):6331–41. [BV]

Tang C., Chehayeb D., Srivastava K., Nemenman I. & Sober S. J. (2014) Millisecond-scale motor encoding in a cortical vocal area. *PLOS Biology* **12**:e1002018. doi:10.1371/journal.pbio.1002018. [rRB]

Teller D. Y. (1984) Linking propositions. *Vision Research* **24**:1233–46. [aRB]

Thompson E., Palacios A. & Varela F. J. (1992) Ways of coloring: Comparative color vision as a case study for cognitive science. *Behavioral and Brain Sciences* **15**:1–26. doi:10.1017/S0140525X00067248. [rRB]

Thompson F. B. (1968) The organization is the information. *American Documentation* **19**:305–08. [aRB]

Thompson S. K., von Kriegstein K., Deane-Pratt A., Marquardt T., Deichmann R., Griffiths T. D. & McAlpine D. (2006) Representation of interaural time delay in the human auditory midbrain. *Nature Neuroscience* **9**:1096–98. [aRB]

Thura D. & Cisek P. (2014) Deliberation and commitment in the premotor and primary motor cortex during dynamic decision making. *Neuron* **81**(6):1401–16. [PC]

Thura D. & Cisek P. (2016) Modulation of premotor and primary motor cortical activity during volitional adjustments of speed-accuracy trade-offs. *Journal of Neuroscience* **36**(3):938–56. [PC]

Todorov E. (2009) Parallels between sensory and motor information processing. In: *The Cognitive Neurosciences* (4th edition), ed. M. S. Gazzaniga, pp. 613–24. MIT Press. [MB]

Tonegawa S., Liu X., Ramirez S. & Redondo R. (2015) Memory engram cells have come of age. *Neuron* **87**:918–31. [aRB]

Trouche S., Perestenko P. V., Ven G. M., van de Bratley C. T., McNamara C. G., Campo-Urriza N., Black S. L., Reijmers L. G. & Dupret D. (2016) Recoding a cocaine-place memory engram to a neutral engram in the hippocampus. *Nature Neuroscience* **19**(4):564–67. [SRS]

Tsai J. J., Koka K. & Tollin D. J. (2010) Varying overall sound intensity to the two ears impacts interaural level difference discrimination thresholds by single neurons in the lateral superior olive. *Journal of Neurophysiology* **103**:875–86. doi:10.1152/jn.00911.2009. [rRB]

Tsotsos J. K. (2011) *A computational perspective on visual attention*. MIT Press. [GNR]

Turing A. M. (1936) On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* **2**(1):230–65. [SH]

Turing A. M. (1950) Computing machinery and intelligence. *Mind* **49**:433–60. [SH]

Turvey M. T. & Fonseca S. T. (2014) The medium of haptic perception: A tensegrity hypothesis. *Journal of Motor Behavior* **46**(3):143–87. [FK]

Tytell E. D., Holmes P. & Cohen A. H. (2011) Spikes alone do not behavior make: Why neuroscience needs biomechanics. *Current Opinion in Neurobiology* **21**(5):816–22. [FK]

Umiltà M. A., Kohler E., Gallese V., Fogassi L., Fadiga L., Keysers C. & Rizzolatti G. (2001) I know what you are doing: A neurophysiological study. *Neuron* **31**(1):155–65. [JB]

Uttal W. R. (1973) *The psychobiology of sensory coding*. Psychology Press. [aRB]

van den Oord A., Dieleman S., Zen H., Simonyan K., Vinyals O., Graves A., Kalchbrenner N., Senior A. & Kavukcuoglu K. J. S. (2016) WaveNet: A generative model for raw audio. arXiv:1609.03499n[cs.SD]. [DB]

van den Oord A., Kalchbrenner N. & Kavukcuoglu K. (2016) Pixel recurrent neural networks. *Proceedings of Machine Learning Research* **48**:1727–36. [DB]

van Gelder T. (1995) What might cognition be, if not computation? *Journal of Philosophy* **92**(7):345–81. [arRB]

van Gelder T. (1998) The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences* **21**(5):615–28. [arRB]

Varela F. G., Maturana H. R. & Uribe R. (1974) Autopoiesis: The organization of living systems, its characterization and a model. *Biosystems* **5**:187–96. doi:10.1016/0303-2647(74)90031-8. [rRB]

Varela F. J., Thompson E. & Rosch E. (1991) *The embodied mind: Cognitive science and human experience*. MIT Press. [XDA, CH]

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L. & Polosukhin I. (2017) Attention is all you need. In: *Advances in neural information processing systems (NIPS 2017)*, ed. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett, pp. 5998–6008. Neural Information Processing Systems Foundation. [AS]

Verschure P. F., Voegtlin T. & Douglas R. J. (2003) Environmentally mediated synergy between perception and behaviour in mobile robots. *Nature* **425**(6958):620. [XDA]

Vilares I., Howard J. D., Fernandes H. L., Gottfried J. A. & Kording K. P. (2012) Differential representations of prior and likelihood uncertainty in the human brain. *Current Biology* **22**:1641–48. [ISJ]

von der Malsburg C. (1999) The what and why of binding: The modeler's perspective. *Neuron* **24**:95–104. [aRB]

von Frisch K. (1967) *The dance-language and orientation of bees*. Harvard University Press. [CRG]

von Frisch K. & Lindauer M. (1954) Himmel und Erde in Konkurrenz bei der Orientierung der Bienen. *Die Naturwissenschaften* **41**(1 June):245–53. [CRG]

von Uexküll J. (1909) *Umwelt und Innenwelt der Tiere*. Springer. Available at: http://archive.org/details/umweltundinnenwe00uexk. [Accessed December 17, 2018.] [aRB]

von Uexküll J. (1926) *Theoretical biology*. Harcourt Brace. [AG-M]

von Uexküll J. (1992) A stroll through the worlds of animals and men: A picture book of invisible worlds. *Semiotica* **89**(4):319–91. [Ld-W]

Wacogne C., Labyt E., van Wassenhove V., Bekinschtein T., Naccache L. & Dehaene S. (2011) Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences* **108**(51):20754–59. [BV]

Walsh D. M. & Ariew A. (1996) A taxonomy of functions. *Canadian Journal of Philosophy* **26**(4):493–514. [DB]

Wang H. & Yang J. (2016) Multiple confounders correction with regularized linear mixed effect models, with application in biological processes. bioRxiv, November. Cold Spring Harbor Laboratory, 089052. doi:10.1101/089052. [RAG]

Wark B., Lundstrom B. N. & Fairhall A. (2007) Sensory adaptation. *Current Opinion in Neurobiology* **17**:423–29. doi:10.1016/j.conb.2007.07.001. [rRB]

Warland D. K., Reinagel P. & Meister M. (1997) Decoding visual information from a population of retinal ganglion cells. *Journal of Neurophysiology* **78**(5):2336–50. [CH]

Weber M. (1978) *Economy and society: An outline of interpretive sociology*. University of California Press. [rRB]

Williamson R. S., Ahrens M. B., Linden J. F. & Sahani M. (2016) Input-specific gain modulation by local sensory context shapes cortical and thalamic responses to complex sounds. *Neuron* **91**(2):467–81. [CH]

Wise S. P. (1985) The primate premotor cortex: Past, present, and preparatory. *Annual Review of Neuroscience* **8**:1–19. [PC]

Wolpaw J. R., Birbaumer N., Heetderks W. J., McFarland D. J., Peckham P. H., Schalk G., Donchin E., Quatrano L. A., Robinson C. J. & Robinson C. J. (2000) Brain-computer interface technology: A review of the first international meeting. *IEEE Transactions on Rehabilitative Engineering* **8**:164–73. [ISJ]

Wolpert D. H. & Macready W. G. (1997) No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* **1**:67–82. [ISJ]

Wood C. C. (2019) The computational stance in biology. *Philosophical Transactions of the Royal Society B Biological Sciences* **374**:20180380. doi:10.1098/rstb.2018.0380. [rRB]

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press

Woolley S. M., Gill P. R., Fremouw T. & Theunissen F. E. (2009) Functional groups in the avian auditory system. *Journal of Neuroscience* **29**(9):2780–93. [CH]

Wurtz R. H. (2009) Recounting the impact of Hubel and Wiesel. *Journal of Physiology* **587**:2817–23. [ISJ]

Yeon J. & Rahnev D. (2019) The nature of the perceptual representation for decision making. bioRxiv:537068 Available at: https://www.biorxiv.org/content/10.1101/537068v1. [Accessed March 5, 2019.] [DR]

Yi K., Wu J., Gan C., Torralba A., Kohli P. & Tenenbaum J. (2018) Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In: *Advances in Neural Information Processing Systems, Volume 31*, ed. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi & R. Garnett, pp. 1031–42. Neural Information Processing Systems Foundation. [JGau]

Yin T. C. & Chan J. C. (1990) Interaural time sensitivity in medial superior olive of cat. *Journal of Neurophysiology* **64**:465–88. [aRB]

Yoo S. B. M. & Hayden B. Y. (2018) Economic choice as an untangling of options into actions. *Neuron* **99**(3):434–47. [PC]

Yoshihara M. & Yoshihara M. (2018) "Necessary and sufficient" in biology is not necessarily necessary – Confusions and erroneous conclusions resulting from misapplied logic in the field of biology, especially neuroscience. *Journal of Neurogenetics* **32**:53–64. doi:10.1080/01677063.2018.1468443. [rRB]

Yost W. A. & Zhong X. (2014) Sound source localization identification accuracy: Bandwidth dependencies. *The Journal of the Acoustical Society of America* **136**:2737–46. doi:10.1121/1.4898045. [rRB]

Young M. P. & Yamane S. (1992) Sparse population coding of faces in the inferotemporal cortex. *Science* **256**:1327–31. [SRL]

Zatorre R. J., Halpern A. R., Perry D. W., Meyer E. & Evans A. C. (1996) Hearing in the mind's ear: A PET investigation of musical imagery and perception. *Journal of Cognitive Neuroscience* **8**(1):29–46. [BV]

Zhurov Y. & Brezina V. (2006) Variability of motor neuron spike timing maintains and shapes contractions of the accessory radula closer muscle of aplysia. *Journal of Neuroscience* **26**:7056–70. doi:10.1523/JNEUROSCI.5277-05.2006. [rRB]

Zylberberg J. (2018) The role of untuned neurons in sensory information coding. Preprint. bioRxiv:134379. Available at: https://doi.org/10.1101/134379. [aRB]

https://doi.org/10.1017/S0140525X19000049 Published online by Cambridge University Press