Neural coding: the bureaucratic model of the brain

Romain Brette

Neural coding: the bureaucratic model of the brain

Romain Brette[1*]
[1] Sorbonne Universités, UPMC Univ Paris 06, INSERM, CNRS, Institut de la Vision, 17 rue Moreau, 75012 Paris, France
* Correspondence: R. Brette, Institut de la Vision, 17, rue Moreau, 75012 Paris, France
romain.brette@inserm.fr
http://romainbrette.fr

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.

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 (R1)? 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 perspective (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 (R2), 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 (R3). 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.

R1. The metaphorical nature of neural codes

R1.1. 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", 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 stakes and needs to be defined. Schultz & Gava propose that neural codes are "the rules governing how sensory events in the world map on to neuronal activity", 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". **Gauthier et al.** claim that "neural codes must implement the empirically validated representations of computational models". 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 et al.** agree that the neural coding discourse often improperly focuses on what the 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". 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", 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 multi-patterned, 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 counter-examples of dominant theories instead of just supporting them". 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", 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 (**Huetz et al.; Gauthier et al.; Birch**), while others regretted my commitment to representations (**Keijzer; Aranyosi; Harnad**). 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 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.". 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.

R1.2. 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 perspective (**Keijzer; Arsiwalla et al.; Gomez-Marin**). 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": 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 the observer to project their own perspective onto the organism (**Gomez-Marin; Cao & Rathkopf**). As **Gomez-Marin** puts it, "a description of what the neuroscientist can do prescribes what the animal must do". 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", and finally "neural codes *must* implement the empirically validated representations of computational models" (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", concluding that these angles must be encoded by the brain.

This methodology follows David Marr's classical three levels of analysis (Marr, 1982): 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. Tony Bell has argued convincingly (Bell, 1999) 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** explains 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.".

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". 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" is 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". 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. Secondly, 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 R3).

R1.3. 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 in order 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". 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, etc. 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 et al.** point out that recent progresses in artificial intelligence have generally ignored coding considerations : "the richest theoretical insights, emerged from studying control, optimization, and learning processes rather than the particularities of representations or codes". 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 motivation when he proposed the efficient coding hypothesis (Barlow, 1961), and presumably what **Harnad** has in mind when he finds it "harmless to call the neural activity along sensory input pathways a "neural code."".

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". Leaving aside the issue that this claim is supported by behavioral rather 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 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, i.e., 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**).

R2. 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", 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 check-list 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.

R2.1. The parking lot

The parking lot of an office building has ten spaces, but there are twelve 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 employee's arrival time that causes its 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". 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 and 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."

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." 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.

R2.2. 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 is consistent with an overall picture of a circular, coupled causal system.". 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.
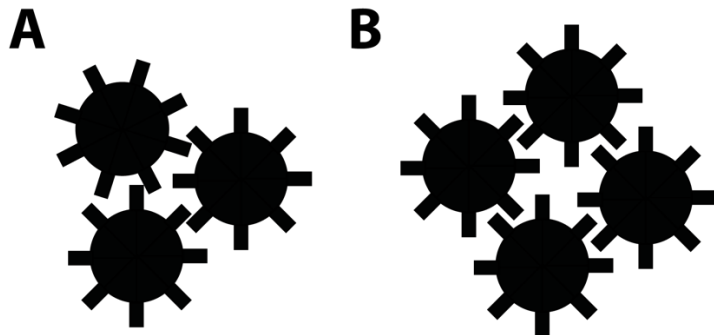

Figure 1. *Two systems of gears with different functionality.*

Consider the system of three gears in Figure 1A. 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 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. 1B) and a dysfunctional (Fig. 1A) 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 bi-partite (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", misses this point.

R2.3. 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), 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". **Gauthier et al.** explicitly consider "brains as representational and computational devices".

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, 1982, Fig. 1-4 p25). 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 C. C., 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 and 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?

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?

R3. Representations

R3.1. Mental representations vs. neural encodings

As many commentators have noted, the neural coding metaphor revolves around a central concept in philosophy of mind: representation (**de-Wit et al.**; **Aranyosi**; **Huetz et al.**; **Lehky & Sereno**; **Mirski & Bickhard**; **Cisek**; **Keijzer**; **Gauthier et al.**; **Deacon & Rączaszek-Leonardi**; **Cao & Rathkopf**; **Jones & Kording**; **Birch**). 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, p50): 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, 1991; Gelder, 1998; Powers, 1973) 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 (IID) 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", 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 IID by subtracting the intensity of the two monaural signals. A few authors noted that those neurons are also sensitive to ITD (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 & Tollin (Benichoux 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". 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", 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 (**Huetz et al.; Gauthier et al.; Birch**). 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 & 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 (**Keijzer; Aranyosi; Harnad**), 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 affordances (Gibson, 1979) 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**) No, because there are ways to conceive these important aspects of representation without neural codes.

R3.2. 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". 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 some thing 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.

R3.3. 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, 1973) sees behavior as the "closed-loop control of what the animal senses" (Arsiwalla et al.).

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 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 turn-over). 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: "Since 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."

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". 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

Andersen F, Anjum RL, Rocca E. 2019. Philosophical bias is the one bias that science cannot avoid. *eLife* **8**:e44929. doi:10.7554/eLife.44929

Anderson ML, Chemero T. 2013. The problem with brain GUTs: Conflation of different senses of "prediction" threatens metaphysical disaster. *Behavioral and Brain Sciences* **36**:204–205. doi:10.1017/S0140525X1200221X

Barlow H. 1961. Possible principles underlying the transformations of sensory messages In: Rosenblith W, editor. Sensory Communication. MIT Press. pp. 217–234.

Bell AJ. 1999. Levels and loops: the future of artificial intelligence and neuroscience. *Philos Trans R Soc Lond B Biol Sci* **354**:2013–2020.

Benichoux V, Rébillat M, Brette R. 2016. On the variation of interaural time differences with frequency. *The Journal of the Acoustical Society of America* **139**:1810–1821. doi:10.1121/1.4944638

Benichoux V, Tollin DJ. 2018. These are not the neurons you are looking for. *eLife* **7**:e39244. doi:10.7554/eLife.39244

Bickhard MH. 2009. The interactivist model. *Synthese* **166**:547–591. doi:10.1007/s11229-008-9375-x

Brette R. 2019. Is coding a relevant metaphor for the brain? *Behav Brain Sci* 1–44. doi:10.1017/S0140525X19000049

Brette R. 2015. Philosophy of the Spike: Rate-Based vs. Spike-Based Theories of the Brain. *Frontiers in Systems Neuroscience* 151. doi:10.3389/fnsys.2015.00151

Brooks RA. 1991. Intelligence without representation. *Artificial Intelligence* **47**:139–159. doi:10.1016/0004-3702(91)90053-M

Chemero A. 2011. Radical Embodied Cognitive Science. MIT Press.

Clark A, Toribio J. 1994. Doing without representing? *Synthese* **101**:401–431. doi:10.1007/BF01063896

Franken TP, Joris PX, Smith PH. 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

Gastinger MJ, Tian N, Horvath T, Marshak DW. 2006. Retinopetal Axons in Mammals: Emphasis on Histamine and Serotonin. *Current Eye Research* **31**:655–667. doi:10.1080/02713680600776119

Gelder T van. 1998. The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences* **21**:615–628.

Gibson JJ. 1979. The Ecological Approach to Visual Perception. Routledge.

Gilbert CD, Li W. 2013. Top-down influences on visual processing. *Nature Reviews Neuroscience* **14**:350–363. doi:10.1038/nrn3476

Gomez-Marin A. 2017. Causal Circuit Explanations of Behavior: Are Necessity and Sufficiency Necessary and Sufficient? Decoding Neural Circuit Structure and Function. Springer, Cham. pp. 283–306. doi:10.1007/978-3-319-57363-2_11

Gordon G. 1994. Workable Gears, Archimedian Solids and Planar Bipartite Graphs. *The American Mathematical Monthly* **101**:527–534. doi:10.2307/2975318

Hofman PM, Van Opstal AJ. 1998. Spectro-temporal factors in two-dimensional human sound localization. *J Acoust Soc Am* **103**:2634–2648.

Joris PX, Yin TC. 1995. Envelope coding in the lateral superior olive. I. Sensitivity to interaural time differences. *J Neurophysiol* **73**:1043–1062.

Laughlin S. 1981. A simple coding procedure enhances a neuron's information capacity. *Z Naturforsch, C, Biosci* **36**:910–912.

Lettvin JY, Maturana HR, McCulloch WS, Pitts WH. 1959. What the Frog's Eye Tells the Frog's Brain. *Proceedings of the IRE* **47**:1940–1951. doi:10.1109/JRPROC.1959.287207

Marr D. 1982. Vision, First Edition. ed. W.H.Freeman & Co Ltd.

Martyushev LM. 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. *Phys Life Rev* **24**:40–41. doi:10.1016/j.plrev.2017.11.010

Maturana HR, Varela FJ. 1973. Autopoiesis and Cognition: The Realization of the Living, 1st edition. ed. Dordrecht, Holland; Boston: D. Reidel Publishing Company.

Populin LC, Yin TC   T. 1998. Behavioral Studies of Sound Localization in the Cat. *J Neurosci* **18**:2147–2160.

Powers WT. 1973. Behavior: The control of perception, Behavior: The control of perception. Oxford, England: Aldine.

Sabin AT, Macpherson EA, Middlebrooks JC. 2005. Human sound localization at near-threshold levels. *Hearing Research* **199**:124–134. doi:10.1016/j.heares.2004.08.001

Sober SJ, Sponberg S, Nemenman I, Ting LH. 2018. Millisecond Spike Timing Codes for Motor Control. *Trends in Neurosciences*, Special Issue: Time in the Brain **41**:644–648. doi:10.1016/j.tins.2018.08.010

Sterling P. 2012. Allostasis: a model of predictive regulation. *Physiol Behav* **106**:5–15. doi:10.1016/j.physbeh.2011.06.004

Tang C, Chehayeb D, Srivastava K, Nemenman I, Sober SJ. 2014. Millisecond-Scale Motor Encoding in a Cortical Vocal Area. *PLOS Biology* **12**:e1002018. doi:10.1371/journal.pbio.1002018

Thompson E, Palacios A, Varela FJ. 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

Tsai JJ, Koka K, Tollin DJ. 2010. Varying overall sound intensity to the two ears impacts interaural level difference discrimination thresholds by single neurons in the lateral superior olive. *J Neurophysiol* **103**:875–886. doi:10.1152/jn.00911.2009

van Gelder T. 1995. What Might Cognition Be, If Not Computation? *The Journal of Philosophy* **92**:345–381. doi:10.2307/2941061

Varela FG, Maturana HR, Uribe R. 1974. Autopoiesis: The organization of living systems, its characterization and a model. *Biosystems* **5**:187–196. doi:10.1016/0303-2647(74)90031-8

Wark B, Lundstrom BN, Fairhall A. 2007. Sensory adaptation. *Current Opinion in Neurobiology*, Sensory systems **17**:423–429. doi:10.1016/j.conb.2007.07.001

Weber M. 1978. Economy and Society: An Outline of Interpretive Sociology. University of California Press.

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

Yoshihara Motojiro, Yoshihara Motoyuki. 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. *J Neurogenet* **32**:53–64. doi:10.1080/01677063.2018.1468443

Yost WA, Zhong X. 2014. Sound source localization identification accuracy: Bandwidth dependencies. *The Journal of the Acoustical Society of America* **136**:2737–2746. doi:10.1121/1.4898045

Zhurov Y, Brezina V. 2006. Variability of Motor Neuron Spike Timing Maintains and Shapes Contractions of the Accessory Radula Closer Muscle of Aplysia. *J Neurosci* **26**:7056–7070. doi:10.1523/JNEUROSCI.5277-05.2006