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

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,