NEWS & VIEWS

doi: 10.30.2017/crockerNSCI425

NEUROSCIENCE

How explicit is the neural code?

The relation of meaning to neural activity is one of the holy grails of neuroscience. Here, Quiroga et al. (2005), is reviewed in the context of the debate surrounding the extent to which the neural code is localized in single neurons, functionally identifiable neural networks, or broadly distributed throughout the brain.

CHRISTOPHER DIAK[1,2]

One of the deepest problems in computational neuroscience is the neurobiological basis of meaning. It is often assumed that the brain computes over discrete functional units at various levels of the nervous system, to create neural representations of objects and events which are integrated into a meaningful experience. At the localist extreme, the activity of single neurons is taken to be a perceptually significant event.[2] On the other end of the spectrum, there are cognitive scientists who believe the brain itself can be modelled as a universal Turing machine, a programmable general purpose computer which computes over perceptual symbols to form mental representations.[8]

Several neurophilosophical problems must be addressed to adequately account for the neurophysiological data which point in a localist direction,[3,6,7,11] namely, the functional unity of high-level cognitive processes such as language generation and use, consciousness, and emotion. Functional brain activity can be characterized at vastly different scales, and the distance between them is daunting; does the brain generate mental representations at different scales using general mechanisms and principles? Is this incompatible with a localist picture? In 1972, Horace Barlow famously articulated 5 dogmas to guide the development of perceptual psychology[2]:

1) behavior depends on cellular level interactions;
2) the sensory system builds maximal representations from minimal components;
3) neurons are selectively and differentially responsive to environmental stimuli;
4) neurons involved in high-level processes cause particular elements of experience;
5) frequency of neural spiking is a plausible mechanism of subjective certainty that features of the objective world exist;

While this list may be somewhat antiquated and vague, forty-five years later the localist picture of the brain hasn't deviated much from the dogma. When Quiroga et al. set out to look for invariant visual representations in neurons they based their approach on the idea that it is conceptually coherent for a single neuron's activity to encode some abstract meaningful information. There are specific prescientific notions of meaning shaping this pursuit. After discussing their findings, we'll deconstruct those notions at some length, and their implications for a localist theory of knowledge. Before we get there, it'll be important to resolve the ambiguity of the term *representation*, as it is used fairly loosely in Quiroga et al. (2005).

A *neural representation* is a specific, physical network of cells (or a single cell) that is theorized to accompany a *mental* or *abstract representation*: the functional units of cognition which may, or may not, have a *meaning*, which is, for the sake of our argument, the referent of the mental representation. The danger of misusing the term should be clear: you can end up with a picture of brain function that is based on a metaphor. For example Quiroga et al. say at the

[1] Program in Neuroscience, Middlebury College, Middlebury, VT, 05753
[2] Center for Cognitive Neuroscience, University of Pennsylvania, Philadelphia PA, 19104

beginning of their paper, that the motivation for pursuing single neuron invariant representations was an unpublished prior result of single units responding to famous individuals.[6] Quiroga et al. hypothesized that "neurons might encode an abstract representation of an individual." Neurons cannot encode *abstract* representations. Neurons are physical, whereas abstract representations are… well, abstract. I'm not arguing for dualism; I *am* saying that if behavior depends on cellular level interactions, the localist picture is *incoherent* because the behavior "recognizing Jennifer Aniston" has no *meaning* without an organism-level description of the behavior; but, by hypothesis, behavior depends upon and knowledge is localized in cellular interactions.

Quiroga et al are not especially at fault in their approach. This is a pervasive source of confusion in modern neuroscience.[3,5,6,7] Unfortunately, this cannot be written off as a mere difference in the kind of questions asked by neurobiology and cognitive science; we are asking the same question: how does the biological entity we call the brain generate mental life? Specifically to Quiroga et al, the question being asked is how this organ generates robust abstract representations of individuals, objects, and landmarks. To address the question, the authors argued for an invariant, sparse, and explicit code on the basis of their electrophysiological data. In this study, Quiroga et al. recorded from a total of 993 units using depth electrodes in eight patients with epilepsy. As only clinical criteria could be used for determining the placement of the electrodes, the data was arbitrarily constrained in some respects. The authors chose the medial temporal lobe (MTL) to investigate because they'd previously characterized neurons which fired selectively in the MTL in response to faces, animals, objects or scenes.[9]

The hypothesis that abstract representations of individuals could be local to MTL cells was well-motivated in the framework the authors were working under: the MTL includes the structures of the entorhinal and perirhinal cortices, the hippocampus and perihippocampal cortex.[9] The entorhinal cortex is the known locus of a spatial map composed of grid cells which activate in response to an animal's location on a vertex of a projected two-dimensional map composed of equilateral triangles[5]; the hippocampus has a well characerized role in episodic and spatial memory formation and consolidation.[9] The MTL is a natural target for researchers looking to understand how the brain implements abstract representations. The challenge for neuroscientists is to get clear about meaning, representation, and the nature of mental events.

What is the idea of meaning implicit in a neural code which is invariant, sparse, and explicit? To use a phrase of Wittgenstein's, we are to imagine that the activation of these neurons is akin to the striking of notes on the keyboard of the imagination. Simply put, this cannot be true.

In Figure 1, Quiroga et al. show the activity of a single unit (a single depth electrode) in response to 30 of 87 images shown to one patient. Of which, every picture of Jennifer Aniston caused at least one spike. The average stimulation was 4.85 spikes (s.d. = 3.59), well above the 5 standard deviation threshold (0.82 spikes in this unit) the authors chose for declaring a statistically significant finding. The responses were observed 300 to 600 ms after stimulus onset. It seems we have our piano key and its note. The authors used a receiver operating framework (ROC), to quantify the degree of invariance (the likelihood of false positives) of the responses. They generated 99 surrogate curves by randomly selecting among the images and calculating a value for each curve by dividing the hit rate (whether or not the unit fired) by the relative number of responses to other pictures. Thus, curves with values of 0.5 would represent firing at chance rate, and curves with values of 0.99 would represent ($P < 0.01$) rate. The area under the curve representing all pictures of Jennifer Aniston is 1.00.

How should we interpet this finding? The authors present similar evidence in Figure 2 for a "Halle Berry neuron" which

responds only to pictures of, drawings of, or the written words "Halle Berry". Again it seems we've found a piano key. To the authors' credit, they explicitly state that their findings should not be viewed as evidence for grandmother cells for three reasons: 1) some units responded to multiple objects, 2) other stimuli could have potentially caused firing, and 3) the speed with which they found 1:1 correspondence between individuals and neural spikes suggest that the cells may represent classes of images.[7]

Again there is a pun being used when the authors speak of representation. There is a saying among cognitive scientists, "no computation without representation," which means that a semantic interpretation must accompany a syntactic operation to count as a computation.[8] It is easy to see how it might apply in the context of single-neuron studies of abstract representation: suppose the localist position that a sparse, invariant, explicit code exists within the brain. Importantly, assume this code is not constructive (i.e., that it doesn't use a combinatorial process for combining metric features of a situation to generate a representation, as Quiroga et al argue against). For the computation resulting in the output "recognize Jennifer Aniston" a neural representation would need to encode not only the semantic information related to Jennifer Aniston, but also the transformation operation used by the brain to turn the electrical stimulation into a perceptual experience. If it didn't, there would be no semantic interpretation, no representation and thus no computation. In other words, the interpreter must be coded into the circuit along with the representational content because otherwise the behavior "recognize Jennifer Aniston" has *no meaning, nothing it refers to,* at the cellular level. Add to this picture the idea that a different neural circuit encodes "recognize Halle Berry" and the the incoherence is manifest. On the localist picture, the computation which generates the semantic interpretation of the neural computation "recognize Halle Berry" would be locally realized in that circuit and would be a different computation than that which gave us "recognize Jennifer Aniston," but if the process of generating a semantic representation is shared between circuits as it would *have to be* if there were an algorithm, a sparse, explicit neural code that were computing over these elements, then positing the existence of an interpreter in each circuit is redundant.

To recapitulate, the localism advocated by Quiroga et al. is a logically incoherent position arising from a misuse or misunderstanding of the term "representation." The position is incoherent because an invariant neural code suggests a common computational system across the brain, but representations are taken to be encoded in neurons or circuits that implement local computations to represent meaning.

What can we hope to learn from studies at the single-neuron level about the representation of meaning? Rather than approach the subject dogmatically, it would make sense to characterize the logic neurons implement to generate their action potentials. In a groundbreaking paper published in *Nature*, Jia et al. (2010) do just this: using a novel approach to visualizing dendritic activity *in vivo*, they demonstrate that neurons which are tuned to a particular stimulus type in the visual cortex compute their firing pattern by integrating spatially over the synaptic activity of the entire dendritic tree.[4] Using a calcium-sensitive dye to track synaptic activity on cortical dendrites, they arrived at the surprising conclusion that cortical dendrites have hotspots which are each tuned for particular stimulus types, and that these hotspots are distributed randomly throughout the dendritic tree.[4] Notice the distinction between the computation described in Jia et al. (2010) and Quiroga et al. (2005). Jia et al. describe a process for computing a neuron's firing rate from its various electrical inputs, and they claim this computation has, as its representation, the *firing rate* of the neuron.[4] Granted, Jia et al. are not asking the question of how this organization translates to meaning, but it does show how an exhaustive representation of the electrical

activity of neurons can be postulated in terms of its action potentials. To get to meaning, consciousness, or any other high-level cognitive process, we can either assume brute identity between firing rates and aspects or elements of meaning, or we have to search for other mechanisms at other levels of the brain. It is the position of the author that this is not only convenient, but necessary, as mental representations and their meanings cannot be reduced to their neural substrate.

In the Chatterjee Laboratory at the Center for Cognitive Neuroscience at the University of Pennsylvania, research questions on meaning, representation, and abstraction are handled more deliberately. A quick review of Chatterjee's publications give a sense of the attention to philosophical, as well as neurological, detail implicit in the research. For example, in Amorapanth et al. (2012), the team investigated the relation between spatial language and abstract spatial representations in subjects with right-hemisphere damage and left-hemisphere damage. 17 patients with left-hemisphere damage (LHD) and 17 patients with right-hemisphere damage (RHD) were given a series of psychological examinations, some visuospatial, some verbal, to tease apart the relation between these two high-level processes. In spite of a positive result, suggesting that LHD patients suffered representational deficits, the team conducted focused residual analyses using voxel-based lesion symptom mapping (VLSM) and discovered an ambiguity between a true representational deficit and expressive deficits.[1] From the above arguments against localism, it may be thought that any attempt at localizing brain activity is misguided. That would be an unwarranted jump, and would be deeply inconsistent with physiological,[3,11] imaging,[1,9] and psychological[1] data suggesting the localization of function across a wide class of brain structures. Amorapanth et al. claim their findings indicate the verbal, conceptual, and perceptual representations in the brain have parallel structures; instead of attempting to reduce these representations to one another, these authors instead provide a guide for developing models of the neural instantiation of abstract representations:

However these abstract representations are neurally represented, we can expect their organizations will be similar.

Rather than attempt to reduce one kind of representation to another, as Quiroga et al have done, we have to ask whether it is plausible for a given unit of some cognitive function to be characterized at a given level. There is no a priori reason why the various functions in the brain have to derive from the same neural code; such a picture would suggest that the brain is literally a universal Turing machine, a radical view which has fallen out of favor in the last two decades due to the rise of the embodied cognition research program[10].

The problem may be the mind as machine metaphor. It is taken for granted that whatever aspect of cognition generates abstract or mental representations, it will be "mechanical"[1,2,5,7,8,9,10,11] because the elementary constituents of the brain are mechanical. It ain't necessarily so. A computational theory of the mind is agnostic as to the nature of the computation. It may well be that *that* nature cannot be represented by our neural code. ■

Christopher Diak is a neuroscience major at Middlebury College and research associate of the Chatterjee Lab at the Center for Cognitive Neuroscience at the University of Pennsylvania.
email: cdiak@middlebury.edu

1. Amorapanth, P., Kranjec, A., Bromberger, B., Lehet, M., Widick, P., Woods, A.J., Kimberg, D.Y. & Chatterjee, A. (2012). Language, perception, and the schematic representation of spatial relations. Brain and Language, 120, 226-236. doi:10.1016/j.bandl.2011.09.007.
2. Barlow, H. Single units and sensation: a neuron doctrine for perception. Perception 1, 371–-394 (1972).
3. Ekstrom, A. D. Kahana M.J., Caplan J.B., Fields, T.A., Isham, E.A., Newoman E.L., Fried, I. Cellular networks underlying human spatial navigation. Nature 425, 184–187 (2003).
4. Jia, H., Rochefort, N.L., Chen, X., Konnerth A. Dendritic organization of sensory input to cortical neurons *in vivo*. Nature. (2010). doi:10.1038/nature08947
5. Hafting T., Fyhn M., Molden S., Moser M., Moser E.I. Microstructure of a spatial map in the entorhinal cortex. Nature 436, 801-806. (2005).
6. Kreiman, G., Koch, C. & Fried, I. Category-specific visual responses of single neurons in the human medial temporal lobe. Nature Neurosci. 3, 946–953 (2000).
7. Quiroga R.Q., Reddy L., Kreiman G., Koch C., Fried I. Invariant visual representation by single neurons in the human brain. Nature Vol 435, 1102-1107 (2005).
8. Rescorla, M. The Computational Theory of Mind. The Stanford Encyclopedia of Philosophy. (2017).
9. Squire, L. R., Stark, C. E. L. & Clark, R. E. The medial temporal lobe. Annu. Rev. Neurosci. 27, 279–306 (2004).
10. Wilson, R.A., Foglia, L. Embodied Cognition. The Stanford Encyclopedia of Philosophy. (2017).
11. Young, M. P. & Yamane, S. Sparse population coding of faces in the inferior temporal cortex. Science 256, 1327–1331 (1992)

Middlebury College Honor Code

I have neither given nor received unauthorized aid on this assignment,
Christopher James Diak
10.30.17