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

Rosa Cao[1] and Charles Rathkopf[2]

[1]Department of Philosophy, Stanford University, Stanford, CA 94305

[2]Institute for Neuroscience and Medicine, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

November 2019

Abstract

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

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

*This is a commentary published in *Behavioral and Brain Sciences* (2019), Volume 42, e221. Commentary on: Brette, R. (2019). Is coding a relevant metaphor for the brain? DOI: https://doi.org/10.1017/S0140525X19001420

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

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

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

It is Brette's second point that we want to focus on. If we insist that neural signals have the same rich semantic properties as conventional human symbols, then indeed there seems to be a puzzle as to how neurons could learn such meanings, given the stark differences between the world as scientists see it and what neurons themselves have access to. But why must we insist on such a demanding notion of code? To require an independent reservoir of meaning, from which we can draw semantic labels to stick on neural signals is not just a neuroscientist's fantasy, but a fantasy tout court. Human language itself has no such reservoir to appeal to – ultimately, our symbols acquire

meaning by virtue of our conventions and practices with respect to them. Insofar as action policies can be established among parts of the brain that need to coordinate their activities with each other and the outside world, why shouldn't some neural signals likewise acquire meanings by virtue of their action policies?

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

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

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

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

Of course human practices with respect to the symbols we exchange with each

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

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

References

Cao, Rosa (2012). "A teleosemantic approach to information in the brain". In: *Biology & Philosophy* 27.1, pp. 49–71.

Dretske, Fred (1994). "If you can't make one, you don't know how it works". In: *Midwest Studies in Philosophy* 19, pp. 468–82.

Fodor, Jerry A. (1990). *A theory of content and other essays*. MIT Press.

Millikan, Ruth Garrett (1984). *Language, thought, and other biological categories: New foundations for realism*. MIT Press.

Neander, Karen (1995). "Misrepresenting & malfunctioning". In: *Philosophical Studies* 79.2, pp. 109–41.

Papineau, David (2003). "Is representation rife?" In: *Ratio* 16.2, pp. 107–23.

Rathkopf, Charles (2017). "Neural information and the problem of objectivity". In: *Biology & Philosophy* 32.3, pp. 321–36.

Shea, Nicholas (2018). *Representation in cognitive science*. Oxford University Press.

Skyrms, Brian (2010). *Signals: Evolution, learning, and information*. Oxford University Press.