How representation works is more important than what representations are*

Shimon Edelman
Dept. of Applied Mathematics and Computer Science
The Weizmann Institute of Science
Rehovot 76100, ISRAEL
edelman@wisdom.weizmann.ac.il

December 1995

Abstract

A theory of representation is incomplete if it states "representations are X" where X can be symbols, cell assemblies, functional states, or the flock of birds from *Theaetetus*, without explaining the nature of the link between the universe of X's and the world. Amit's thesis, equating representations with reverberations in Hebbian cell assemblies, will only be considered a solution to the problem of representation when it is complemented by a theory of how a reverberation in the brain can be a representation **of** anything.

It is possible (and, according to some, necessary) to distinguish two problems about mental representation (Cummins, 1989). The first of these, the Problem of Representations, is basically empirical, and has to do with the nature of the representations employed by a given cognitive system (such as the brain of a monkey). Amit's thesis, equating long-term memory representations in the monkey visual system with reverberations in certain cell assemblies in the inferotemporal (IT) cortex, outlines a possible answer to the Problem of Representations, mobilizing in support of the proposal an impressive battery of data from neurophysiology and from the mathematics of attractor systems. The target article does not, however, take a stand on the second problem, called by Cummins the Problem of *Representation* (singular). Here, the central question is how, in principle, can a mental state, as realized, e.g., by a reverberation in IT, *refer* to anything at all in the world (that is, what makes a given brain event the representation of, say, apple).

Let us consider the comparative value of theories addressing one but not the other of the two aspects of the problem of representation mentioned above. Note first that one can formulate a

*A commentary on D. Amit, "The Hebbian paradigm reintegrated: Local reverberations as internal representations", *Behavioral and Brain Sciences* **18**(4), 630-631, December 1995.

useful (i.e., predictive) theory of mental processes while treating the individual representations as black-box entities causally related to events in the world (and to each other). In the philosophy of mind, for instance, some of the doctrines that account for behavior in terms of the agent's beliefs and desires treat internal representations as such unanalyzable entities. The great practical value of folk psychology as an explanatory and predictive tool attests to the possibility of a methodological separation between the questions of what representations are and what they are good for. In other words, a solution to the psychophysical problem — getting meanings into and out of the head — can stand on its own, no matter what is, precisely, the meaning of 'meaning.'[1]

Whereas solving the psychophysical problem would constitute a most significant advance on the way to understanding Representation, unraveling the mechanism (such as dynamics of cell assemblies) whereby individual representations may be realized is, at best, a small step (albeit in the right direction). To mix a Fodoresque metaphor with a simile borrowed from Harnad's paper on symbol grounding, seeking out and cataloging representations whose causal relations with the world are unknown or ill-understood amounts to trying to learn Mentalese from a Mentalese/Mentalese dictionary alone. Granted, a dictionary of representations is better than nothing, but without grounding at least some of the entries in reality it is difficult to make progress (Harnad, 1990).[2]

It is interesting to note that experimental neurobiologists appear to take the psychophysical aspect of the problem of representation much more seriously than some theoreticians. One example in support of this observation can be found in (Sakai et al., 1994), where the researchers describe an ingenious method of making sure that the stimulus associated with the response of a certain unit in monkey IT cortex is indeed the optimal one for that unit. In one of the experiments there, the stimulus shape was varied parametrically, leading invariably to a decrease in the unit's response. Under certain conditions on the shape space, this may be interpreted as evidence in support of the original stimulus being the optimal one (note that such evidence, if corroborated, would constitute a solution to the psychophysical problem for that particular neuron).

Another example of progress in this direction is the ground-breaking technique of stimulus reduction, developed by Tanaka and his group (Fujita et al., 1992; Tanaka, 1992). Once identified as effective for a given neuron in IT, a stimulus shape in a typical experiment undergoes successive simplification, until it no longer elicits a response from that neuron. The last stimulus in the reduction sequence that is still effective is the elementary feature that lies at the intersection between the set of features present in the original stimulus and the shape-space "receptive field" of the neuron. The representation inherent in the firing of this neuron is thus grounded in the external world. Note that by narrowing down the range of stimuli to which a particular representational event may refer, the reduction technique offers a partial relief, at a reductionist neurophysiological

[1] In fact, meaning does not have to be in the head at all (Putnam 1988, p.73), provided that whatever *is* in the head obeys a well-defined causal relationship with what is "out there" in the world (Locke, 1690; Edelman, 1995).

[2] As an edifying example, consider Ijon Tichy's attempt to learn the meaning of the Ardrite word "scrupt," and its consequences, as recounted in the fourteenth voyage of the "Star Diaries" (Lem 1985, p.103).

level, of the predicament of indeterminacy of radical translation, illustrated by Quine's celebrated Gavagai example (Quine, 1960).

Stretching somewhat the sense of "footnote," one may consider this emerging neurophilosophy of representation as another footnote to Plato. In *Theaetetus*, Socrates offers (and criticizes) a number of metaphors for the concept of knowledge, one of which is a flock of birds within the cage of one's mind. To solve the problem of representation, it is not enough to understand the dynamics of flocking: one must also find out how to govern the flight of the birds from without the cage.

References

Cummins, R. (1989). *Meaning and mental representation.* MIT Press, Cambridge, MA.

Edelman, S. (1995). Representation, Similarity, and the Chorus of Prototypes. *Minds and Machines*, 5:–.

Fujita, I., Tanaka, K., Ito, M., and Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, 360:343–346.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

Lem, S. (1985). *Star Diaries.* Harcourt Brace Jovanovich, New York.

Locke, J. (1690). *An essay concerning human understanding.* The Internet. available electronically at URL gopher://gopher.vt.edu:10010/02/116/3.

Plato (360). *Theaetetus.* The Internet. translated by B. Jowett; available electronically at URL gopher://gopher.vt.edu:10010/02/131/23.

Putnam, H. (1988). *Representation and reality.* MIT Press, Cambridge, MA.

Quine, W. V. O. (1960). *Word and object.* MIT Press, Cambridge, MA.

Sakai, K., Naya, Y., and Miyashita, Y. (1994). Neuronal tuning and associative mechanisms in form representation. *Learning and Memory*, 1:83–105.

Tanaka, K. (1992). Inferotemporal cortex and higher visual functions. *Current Opinion in Neurobiology*, 2:502–505.