Comment on Ryder's SINBAD Neurosemantics: Is Teleofunction Isomorphism the Way to Understand Representations?

MARIUS USHER

Abstract: The merit of the SINBAD model is to provide an explicit mechanism showing how the cortex may come to develop detectors responding to correlated properties and therefore corresponding to the sources of these correlations. Here I argue that, contrary to the article, SINBAD neurosemantics does not need to rely on teleofunctions to solve the problem of misrepresentation. A number of difficulties for the teleofunction theories of content are reviewed and an alternative theory based on categorization performance and statistical relations is argued to provide a better account and to come closer to the practice in neuroscience and to powerful intuitions on swampkinds and on broad/narrow content.

The SINBAD model is useful in showing how a neural-type model is able to bootstrap itself, learning to develop specific detectors that respond to correlated properties. As mental representations mediate our ability to categorize and identify substances (natural kinds, individuals, etc; Millikan, 1998), which are characterized by correlated properties, the model provides an excellent illustration of how the brain *may* come to develop mental representations. At present, it is not obvious that the specific SINBAD-algorithm is the one utilized by the cortex. Although evidence for backpropagating action potentials into the dendrites exists, the SINBAD-algorithm requires a type of non-local information (dividing by the number of dendrites) that may be difficult to realise in biological neurons, especially if the number of dendrites changes, due to growth or degeneration. Furthermore, the model needs to be tested for its ability to learn categories (corresponding to real or natural kinds) that form the basis of human representation. Ideally, the network learning profile should be shown to mimic that of infants and its learning power should be tested relative to experimental data. Special attention should be given to addressing human and animal limitations in category learning: an algorithm that can learn categories characterized by functions that are too complex for humans and animals is unlikely to offer an adequate mechanism for cortical representations. (I doubt, for example, that human observers can predict the solar planetary system from the mere observation of the planets' trajectories, i.e., without computers and

I want to thank Chris Eliasmith and Nick Zangwill for a critical reading of the commentary and helpful discussions.

Address for correspondence: Department of Psychology, Birkbeck College, Malet Street, London WC1E 7HX, England.
Email: m.usher@bbk.ac.uk

Mind & Language, Vol. 19 No. 2 April 2004, pp. 241–248.

knowledge of the laws of gravity). Finally, it will be important to see how the model achieves the tradeoff between complexity and pattern completion: learning complex patterns while being able to perform under a considerable amount of missing information (as it is geared towards complex logical functios, SINBAD may have more difficulties than simpler Hopfield-type models with pattern completion). These qualifications notwithstanding, the value of the SINBAD model is that it offers an illustration of how a neural cortical mechanism that learns to develop detectors for correlated properties could function. Even if the precise details of the actual cortical mechanism may differ (networks or cell-assemblies versus single pyramidal neurons, etc) the principles may be similar. This illustration is therefore extremely useful in the attempt to develop a theory of *neurosemantics*. In the rest of this commentary, I will focus on the way SINBAD attempts to achieve this.

The SINBAD-neurosemantics theory is based on the idea of *teleofunction isomorphism*. First, SINBAD learns to develop detectors that respond to correlated properties and which become 'standups' for sources of correlations. Second, it learns internal models, isomorphic with the environmental system it interacts with (e.g., the solar planetary system), and it is this property that grounds its neurosemantics. Yet isomorphism (which is a type of 'similarity' relationship) is known to be insufficient for grounding representation (due to a variety of problems mentioned in the article). 'Not all is lost' and so teleofunctions are imported. By analogy with designer based representations (e.g., speedometers), it is suggested that natural representations obtained in SINBAD represent because they are isomorphic with what they are *supposed*, by evolutionary selection, to be isomorphic to; they have this function because this isomorphism provides a predictive skill that was instrumental in the survival of the organisms in which SINBAD networks evolved throughout the species' ancestry.

I believe that the appeal to designer-representation in explaining natural representation is counterproductive and that the quite fashionable reliance on evolutionary selection (teleosemantics) is much overrated. One of the main motivations of this theory is to find a way to fix the content of mental representations, while accounting for the problem of misrepresentation that challenges the causal representation theories (e.g. Stampe, 1977; Dretske, 1983). In these theories, a mental symbol represents the object that caused its tokening. The problem of misrepresentation in causal theories arises because misrepresentations are also tokened by their causes, although they do *not* represent them (Fodor, 1984). Teleosemantics (of the type labeled by Karen Neander (1995), as the High-Church version),[1] for example, is supposed to explain why the frog visual system represents flies and not any dark moving dots, although the frog reacts to any dark moving

[1] All the objections to teleosemantic theories presented here are directed to the High-Church version. I do not argue against the more modest, Low-Church version, which attempts to account for a simpler type of content that corresponds to innate abilities, such as visual textures or direction of motion.

dots in its environments (including flies). According to teleosemantics the frog visual neurons represent flies because it was signaling flies that they were *selected for* by the *consumer* system that makes use of the signal. Being-selected-for is explained in term of history: it was the encounter with flies that made the frog ancestors develop the response to dark moving dots (e.g., Millikan, 1991; Neander, 1995). To summarize, teleosemantics assumes that the content of representations is specified by the environmental feature which was historically responsible for the consumer of the representation performing its teleofunction, so misrepresentations are due to a *malfunctioning* of this system.

Numerous problems that plague the teleosemantic theories of content have been widely discussed by Jerry Fodor (1990; 1991) and others (e.g. Agar, 1993; Antony, 1996; Neander, 1995; Pietroski, 1992; Rey, 2002). It is beyond the scope of this commentary to review this vast literature, however, a few lines may be in order. The first and, probably, the most serious problem is that even if teleosemantics accounts for a type of primitive content (called by Fred Dretske 'modest', as opposed to 'inflated' content), corresponding to innate perceptual representations (such as visual textures or motion direction), no account has yet been given for the content of learned concepts of actual entities (e.g., 'football', 'basketball') or nonexistent ones (e.g., God, angel;[2] it may be hoped that SINBAD can help to fill this gap, but see discussion below.) Second, Jerry Fodor has argued that purely historical accounts (such as those based on evolutionary selection) suffer from intrinsic indeterminacies. For example, he claimed (Fodor, 1990; 1991) that history alone cannot take advantage of the difference between coextensive attributes (such as 'fly' and 'dark-moving-dot', if in the frog's environment most dark-moving-dots are flies) and therefore it cannot account for the statement that the frog neurons are selected for flies and not for mere dark-moving-dots.[3] It is possible to counterargue that history contains not only events described by properties (which may be coexistent) but also causal relations between such properties (perhaps in a non-Humean sense) and that when two properties coexist, evolutionary selection takes place in favour of the causally efficient one (Neander, 1995). This, however, together with the emphasis on the consumer of the representation (Millikan 1991), does not seem to give the desired content for the frog visual neurons. While being a fly (and not a dark moving dot) is the efficient property for the consumer of the frog visual representation (helping the frog's survival), it is not clear why one should stop there, rather than at the property of being a nutrient (Agar, 1993). After all, 'being a fly' is as much of an accidental inefficient property as 'being a dark-moving-dot'; it is being a nutrient that counts as the efficient causal property (and therefore frogs don't represent flies, after all).[4]

---

[2] According to teleosemantics, if the latter have any content at all, this is related to 'keeping people happy/obedient' (Rey, 2002).

[3] As Fodor puts it, Darwin cares how many flies the frog eats and not how you describe them. He further argues that ones needs counterfactuals to distinguish among coextensive properties and that this goes beyond what a purely historical theory should employ (Fodor, 1991).

[4] This underlies a basic problem of the High Church teleosemantics. Even if the frog discriminates between two types of food, say flies and mosquitoes, as long as their representations have the same function for the consumer of these representation (eat them), the two can not be distinguished.

Similarly, the Teleosemantics theory (of the High-Church type) seems to lead to implausible conclusions, such as the statement that 'a male hoverfly misrepresents if he chases an infertile female' (Neander, 1995, p. 127; see also Pietroski, 1992 and Zangwill, 2003, for other examples showing how Teleosemantics gets its content wrong). Third, it is not clear that any practical procedure exists to determine the teleofunction in situations that include, for example, repression. In his reply to Millikan (1991), Fodor asks if the function is 'to token true beliefs (so repression is a malfunction) or is it to token true beliefs *excepting those true beliefs which it is supposed to repress*, in which case repressing those true beliefs isn't a malfunction after all' (p. 294).

There are attempts to answer some of these difficulties. Nevertheless, I estimate that that one of the main drives for teleosemantics is not its impressive success in determining the content of learned representations, but rather a feeling that it is the only game in town providing a type of normativity, which, in turn, is viewed as necessary for semantic evaluations. Accepting this is, however, not without a price. For example it leads one to assert that swampmen (creatures that are accidentally created out of molecules in a swamp, with body and brain structures just like ours but without our evolutionary history) have no beliefs or desires (Millikan, 1996; Neander, 1996).[5] The claim that content cannot be determined without norms or teleofunction is, in its turn, based on two non-proved assertions: i) that content cannot not be determined by statistical dispositions and ii) that misrepresentations are *always* due to 'malfunctioning'. I will try to argue that both of these assumptions are incorrect, by highlighting a recent causal-statistical theory of content and arguing that it can answer these challenges. Before I do so, I examine the way in which SINBAD is formulated in relation to teleofunctions and I will argue that SINBAD does not need teleofunctions and that it could work better within a statistical causal framework.

Dan Ryder argues that SINBAD neurosemantics is not subject to the indeterminacy criticism made by Fodor on the other teleosemantic theories. This is because, unlike in those, in SINBAD specific contents are not fixed by evolution, since the 'etiological account of representational status and the psychosemantics (the specification of content for particular representations) appear to come apart'. Unfortunately however, this fact also makes the biofunctionalism totally redundant to the theory, as I will try to explain below. First remember that fixing specific contents was exactly what biofunctionalism was supposed to buy us, in return for numerous concerns (described above). Second, the two components of SINBAD, the evolution that confers to SINBAD the function of prediction and the mechanism that fixes specific contents, are totally unrelated, and the link between the two appears superfluous. We are told for example that a SINBAD cell may come to respond to a source of

[5] I would even say that it leads to the counterintuitive idea that the only way to mean something is by performing a function, even when the function is beyond the agent's understanding and is totally irrelevant to her present life and activity; the more remote the function the more 'repugnant' this idea appears to the agent (see, e.g., Kurt Vonnegut's, 'Sirens of Titan' for a literary illustration).

correlation whose 'explanatory source' is kind-K and therefore 'we are justified in asserting that the cell is supposed to represent K. This is because the cell is supposed to yield reliable predictions and its corresponding specifically to kind-K is the only reason the cell yields reliable predictions'. But what exactly does this mean? As far as SINBAD is concerned, the same cell could have come to correspond (depending on random pre-learned synaptic connections) to kind-L instead. If that was the case, the cell would still predict. It would predict, however, L-kinds rather than K-kinds. So how can its corresponding to kind-K be the *only reason* the cell yields reliable predictions? Is it not more plausible to conclude that SINBAD cells simply correspond to items that they causally interacted with in their learning history? By coming to respond to K-kinds they became K-representations. True, this confers a predictive skill (and perhaps other skills, such as memory), but this skill is not what makes the cell represent K *rather* than L; the skill was there even if the cell came to respond to L.

Recent theories of natural content based on statistical causal-information theory have shown a way to solve the problem of misrepresentation in causal theories (Eliasmith, 2000; Usher, 2001; see also Dretske, 1983 for a precursor). Unlike teleosemantics, which sees misrepresentations as originating in a malfunctioning of the information processing system, on the statistical approach, misrepresentations are part of the 'normal' function of probabilistic categorisation, grounded in the 'normal' mode of noisy brain processing.[6] According to such theories, causation is a necessary but not a sufficient condition for a mental token to represent an item (or substance). The other necessary condition is that the mental token is statistically correlated in an *optimal* way (as measured by Shannon mutual information) with the item it represents.[7] Since, occasionally, an item can token a mental symbol that corresponds (in the statistical sense above) to another item, the problem of misrepresentation is solved. Moreover, the theory demonstrates that when one correctly uses the Shannon mutual information, situations where, due to high expectation of an item, the tokening of its representation is frequent while its actual frequency is low (e.g., 'danger' for a vulnerable animal that makes lots of false alarms)[8] can be easily dealt with. [The trick is to examine conditional

---

[6] Recent studies indicate that noisy processing is a basic principle of neural computation (Eliasmith, 2000b; Gold and Shadlen, 2002) and that sometimes increasing the noise level is advantageous to performance (especially when a speeded response is needed); this is called *stochastic resonance* (Usher and Feingold, 2001).

[7] Optimal implies a contrastive relation: the mental token represents the item with which its statistical measure of mutual-information is largest (Eliasmith, 2000; Usher, 2001).

[8] Such situations are part of a set of non-proven objections typically made against statistical theories of content (e.g., Millikan, 1989). Other objections include the fact that it is practically impossible to count the counterfactual situations needed for estimating wide-statistical dispositions (Millikan, 1996). Practically, this challenge can be dealt with, as neuroscientists typically do when they use statistical measures of mutual information to determine what neurons represent (Bialek *et al*., 1991). Notice, that all that is needed is an ordinal relationship between the corresponding dispositions and therefore counting all the counterfactuals is not needed; in practice one uses a neutral context, assuming that contexts (counterfactuals) that favour one option or the other balance each other.

probabilities (and not raw correlations), conditioning on the object when contrasting different objects for one mental state, and conditioning on the mental state when contrasting different mental states for one object (and not the other way around); see Usher, 2001, Equations 2–6, for details.]

I propose that the SINBAD neurosemantics can be comfortably formulated within such a framework. After all, SINBAD cells come to respond to items in their environment and the relation between stimulus and response is statistical. Its cells may therefore represent items (kinds), which they statistically correlate with in the sense of transmitted information; when the K-cell fires, one can infer that it is most likely that kind-K (rather than any other kind known by the network) is present. Moreover, a model such as SINBAD provides some additional features that a representation system should posses. First, it contains information about the (correlated) features that characterize the kind (this is in fact its central mechanism). Second, it can generate an internal map of the relations between various kinds, corresponding to inter-conceptual knowledge. Whether such internal maps are necessary (or only secondary) for defining the content of a representation token is debatable. As argued by Fodor and Lepore (1992), the knowledge of conceptual inter-relations changes substantially during the life of individuals, while the content of the mental symbols is thought to stay fixed. Nevertheless, the ability to develop such mental models is definitely an important property (even if not a defining one) of conceptual representations and is necessary for understanding the processes that operate on them.

Let us examine a few contrasts between such a causal information theory of neurosemantics and the normative-etiological one. While in the former, content is mainly specified by the agents' categorization dispositions, in the latter it is specified by its history. Surely, most often, the two go together. Things we experienced and interacted with are most often the ones we know to categorize and identify best now. Divergences between these theories can, however, be imagined. Consider first a situation, where an animal was trained for a prolonged period to respond to a specific category (e.g. categorize pictures of cats and dogs by making specific button responses) and that a cortical cell is found that learned to respond to cat-stimuli. Following this prolonged training the task is changed and now the animal is required to respond to other aspects of the stimulus information, which are orthogonal with the ones it responded first (e.g. animals with long legs). In a set of experiments using such a procedure, Freedman *et al.* (2001) reported neurons in the pre-frontal cortex of monkeys that change their response profile. Whereas at the beginning they responded to cats now they respond to animals with long legs. According to the etiological theory, the cells should probably continue to represent the previous property (cats) since it was this property that it was historically linked with most. Within the causal-information theory, the cell represents instead the item it now stands in correlation with (animals with long legs).[9] (Notice that if the historical account says that now the

[9] Other cells maintain their response profiles. Such cells, however, do not dissociate the two theories, since their history and actual performance profiles overlap.

cell represents 'animals with long legs' because it interacted with such stimuli last, it implicitly admits that the content is fixed with reference to performance; the cell represents these items not because they were the last the agent interacted with but because the cell now responds to them more vigorously than to 'cats').

Another divergence between the two theories appears when considering situations related to broad and narrow content. While the etiological theory insists that twin-earth people who utter the word 'water' do not mean what we mean by 'water', and that swampmen do not mean anything by 'water', the causal-information theory holds that, as long as they know how to categorize and identify water (as well as we do) they mean the same as we do.[10] Despite the presumed consensus for broad content, I find the arguments put forward unconvincing (see also Rey, 2002). For all we know, it is possible that the water structure in Europe and America differs in some elementary particle that makes no difference to any of its properties; we may all then be twin-earth people and, according to the etiological theory, there would be no common content to our representations. Perhaps the time has come to challenge this consensus and support swampmen liberation (Antony, 1996).

Department of Psychology
Birkbeck College

References

Agar, N. 1993: What do frogs really believe? *Australasian Journal of Philosophy*, 93, 337–372.

Antony, L. 1996: Equal right for swamp-persons. *Mind & Language*, 11, 70–75.

Bialek, W., Rieke, F., Van Steveninck, R.R.D. and Warland, D. 1991: Reading a neural code. *Science*, 202, 1854–1857.

Dretske, F. 1983: The epistemology of belief. *Synthese*, 55, 3–19.

Eliasmith, C. 2000: How neurons mean: A neurocomputational theory of representational content. Ph.D. Thesis. Washington University, St. Louis (www.arts.uwaterloo.ca/celiasmi/publications.html)

Eliasmith, C. 2000b: Is the brain analog or digital? The solution and its consequences for cognitive science. *Cognitive Science Quarterly*, 1, 147–170.

Freedman, D.J., Risenhuber, M., Poggio, T. and Miller E. 2001: Categorical representations of visual stimuli in the primate prefrontal cortex. *Science*, 291, 312–316.

[10] In causal-information theory of the type presented in (Usher, 2001; see also, Dretske, 1983; Fodor, 1984), the representation relation supports counterfactuals. Thus, even if water was not the actual cause of 'water' for the twinearth person, the mere possibility of it causing 'water' includes it in the content of 'water'. This content therefore includes all the perceptually indistinguishable objects that can cause the agent to token 'water' more than other mental tokens; the agent has a blended representation for water and twinwater, as she may have for beech/elm, if she she is not a tree-expert.

Fodor, J. 1984: Semantics, Wisconsin style. *Synthese*, 59, 231–250.

Fodor J. 1990: *A Theory of Content and Other Essays*. Cambridge, Mass: MIT Press.

Fodor 1991: Reply to Millikan. In B. Lower and G. Rey (eds.), *Meaning in Mind: Fodor and his Critics*. Oxford, UK: Blackwell (p. 293–296).

Fodor, J and Lepore, E. 1992: Holism: A Shopper's Guide. Oxford: Blackwell.

Gold, J.I. and Shadlen M.N. 2002: Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward, *Neuron*, 36, 299–308.

Millikan, R.G. 1989: Biosemantics. *Journal of Philosophy*, 86, 281–297.

Millikan, R.G. 1991: Speaking Up for Darwin. In B. Lower and G. Rey (eds.), *Meaning in mind: Fodor and his critics*. Oxford, UK: Blackwell.

Millikan, R.G. 1996: On swampkinds. *Mind & Language*, 11, 103–117.

Millikan, R.G. 1998: A common structure for concepts of individuals, stuffs, and real kinds; more mama, more milk and more mouse. *Behavioral and Brain Sciences*, 9, 55–100.

Neander, K. 1995: Misrepresenting and malfunctioning. *Philosophical Studies*, 79, 109–141.

Neander, K. 1996: Swampman meets swampcow. *Mind & Language*, 11, 118–129.

Pietroski, P. 1992: Intentionality and teleological error. *Pacific Philosophical Quarterly*, 73, 267–282.

Stampe, D. 1977: Towards a causal theory of linguistic representation. In P. French, D. Euhling and H. Wettstein (eds.), *Midwest Studies in Philosophy*, 2, 42–63. Minneapolis: University of Minneapolis Press.

Rey, G. 2002: Compromising externalism on behalf of [Angels] and the White Queen. Paper presented at Pacific APA, March (2002).

Usher M. 2001: A statistical referential theory of content: using information theory to account for misrepresentation. *Mind & Language*, 16, 311–334.

Usher M., and Feingold M. 2000: Stochastic resonance in the speed of memory retrieval, *Biological Cybernetics*, 83, L011–L016.

Zangwill, N. 2003: Millikan on direction of fit (in preparation).