# “On Facing up to the Semantic Challenge”

## *1. INTRODUCTION*

Rick Grush recently wrote a paper  on the theoretical constraints of representational accounts in neuroscience (Grush, 2001). In it he presents computational neuroscience with the following challenge: How to distinguish between computation - understood as computational processing of ”genuinely semantic” information - and any other complex causal process, merely governed by a computationally tractable rule?

Obviously, there are countless physical systems that are not computers but in which there nevertheless occur state transitions that can be modeled or simulated computationally. How exactly is one then to distinguish computational systems from computable systems (i.e. systems that *perform* computations form systems that can be *described* by computations)? The natural answer  would be to invoke a semantic notion of computation. The semantic notion of computation holds that computations (performed by the system) are information processing operations over semantically evaluable “mental” entities, contents (here understood as abstract objects), represented in a physical medium (Piccinini 2004). This of course charges one with providing theory of mental content appropriate for the purposes of computational neuroscience, neurosemantics.

Indeed, Grush frames the problem (and his own solution) in terms of semantics, specifically, in terms of a distinction between what he calls a-semantics and e-semantics (a-semantics is isomorphism between the causal neural processes and some abstract algorithm, e-semantics is isomorphism between the causal neural process and the physical causal processes of the environment). He argues that recent computational neuroscience treats computation and representation a-semantically, but that this is inadequate and should be replaced with a more genuinely semantic notion of computation and representation, e-semantics.

In this paper I discuss Grush's challenge and his solution in terms of a slightly different distinction between what I call "horizontal" and "vertical" approaches to assigning a semantics to neural activity. The main point I will argue for is that Grush's proposed alternative to a-semantics (which, in the terminology introduced here, is vertical), his e-semantics, is a version of horizontal content-assignment, but that what one needs to fully address the semantic challenge of computational neuroscience, should be a vertical semantics.  However, what one really needs is a "top-down vertical" semantics - call it c-semantics - rather than  the sort "bottom-up" vertical semantics such as the a-semantics rightly criticized by Grush.

My negative argument against e-semantics and my positive argument for a c-semantic construal of computation and representation is based on considerations concerning the poverty of the stimulus hypothesis and the veridicality assumption (Akins, 1996) prevalent in contemporary neurosemantics and computational neuroscience

## *2. NEUROSCIENCE AND COMPUTATION: THE SEMANTIC CHALLENGE*

There are many types of automata that can compute information, i.e. implement algorithmic computations, by their state transitions. Computational neuroscience is founded on the computational hypothesis of the mind/brain: that complex neural systems are such automata, and that understanding this property of the brain is useful for accounting for much of the intelligence we find in the behavior of organisms.

In a *locus classicus* of the approach, Churchland, Koch & Sejnowski (1990) define the assumption thus (cf. Churchland & Sejnowski, 1992 and Cummins (1989)):

(1) "In a most general sense we can consider a physical system as a computational system just in case there is an appropriate (revealing) mapping between some algorithm and associated physical variables. More exactly, a

> physical system computes a function $f(x)$ when there is (1) a mapping between the system's physical inputs and $x$, (2) a mapping between the system's physical output and $y$, such that (3) $f(x) = y$"

According to definition (1) *all it takes* for a system to compute a function $f$ – all that is required to make the system "a computer" – is for its causal processes to be appropriately equivalent to some algorithm in this sense. Computationality is here clearly identified with computability: if the state transitions conform to some regularity that can be formally captured by function $f$, then that causal process computes the function $f$.

The trouble with this account is basically that computable state transitions occur in countless systems that are *not* computers. Indeed, any computable causal process trivially realizes any algorithm describing its behavior reasonably accurately[1]. What, then, beyond (assumed) computability makes neural systems computational in the sense computational neuroscience is interested in?

The natural response is to fall back on the semantic notion of computation (as defined in Piccinini 2004a, 2004b), and insist that the difference between genuine computation and mere computability is that in semantic computation the entities which computational relations are defined over are (mental) representations, i.e. entities which are semantically related to each other and/or the external world[2]. The semantic notion of computation and representation thus prevents our notion of computation from becoming vacuous.

This is the solution discussed by Grush, and as far as I can see endorsed by many computational neuroscientists as well. For example, Sejnowski et al. (1988) state that:

---

[1] As also discussed, for example, by Putnam and Searle (for historical references see Piccinini, 2004ab).

[2] I.e. an account where the notion of mental representation or semantic information is taken as more fundamental, and computation defined derivatively, rather than the other way round. Computation is to be defined, in the first instance, with explicit reference to mental entities (representations) and operations which operate on the contents of said representations (e.g. Fodor (1981, 1998)). In short: the semantics of a computation are essential to the identity of that computation, and content is constitutive of a computation in that changing the contents defines a different computation, whereas changing the vehicles only makes a difference at the level of representation/algorithm. In externalist or "broad" semantics with which I will be concerned with in this paper, facts about reference are in turn constitutive of content, Dretske (1981,1988), Fodor (1990), Millikan (1989).

> "Mechanical and causal explanations of chemical and electrical signals in the brain are different from computational explanations. The chief difference is that a computational explanation refers to the information content of the physical signals and how they are used to accomplish a task."

Assuming the semantic notion of computation we can see why the solar system, for example (Grush's example), does not "compute" anything in accordance with the laws of Kepler, Newton or Einstein. (Some of) its state transitions may be *computable* in the sense of being *governed by rules* that can be framed in computational terms. The system can be simulated computationally but is not *computational*, in the sense of "performing" those computations in any substantive sense.

Not all computational modeling of neural processes counts as computational neuroscience, then. It is quite possible, and indeed common, to model *physiological mechanisms*, rather than mechanisms at "the representational level" of the brain.

This, of course, means one needs an explicit and non-circular definition of representation and semantic information. To do computational neuroscience in the full sense one needs to be "assigning semantics" to the internal states of the brain that your computational account of the mind and behavior rests on[3].

Yet, looking at one example of contemporary neuroscience (Koch, 1990), Grush (2001) observes that:

> "The [computational account] involving the bullfrog ganglion cell appears to involve no more than would be involved in any example of computer-simulation-cum-experimental-testing endeavor, of the sort familiar in computational physics, economics, meteorology, and dozens of other areas." Grush (2001, p.162).

[3] With human-built models we can ,of course, refer to the designer's intentions and interpretations, a course that is not available in naturalizing semantics of brain states.

The computational description in question involves nothing in the way of a genuinely semantic account of the signals and signal processing operations discovered in the intricate physiology of the neural tissue, as required by the semantic view of computation.

The upshot of Grush’s analyses is that computational neuroscience is in dire need of a principled and workable notion of representation, and the neuroscience itself does not provide an account of what it is to ”semantically compute” something, it presupposes one; hence “the semantic challenge”.

## *2.1. Grush’s Solution: a-semantics and e-semantics*

Grush suggests we should think of the semantic challenge in terms of a-semantics, isomorphism to an algorithm, and e-semantics, isomorphism between environmental variables and the variables of an internal system used in the brain to ”stand in for” the environmental system. In a-semantic, quasi-computational, neuroscience the computational account presents the system as corresponding to, “representing”, any algorithm that it can (“revealingly”) be considered an instantiation of.

Grush’s contention is that what definition (1) above captures, is merely the “a-semantics” of a system, when what you need to do genuine computational neuroscience in the full-blooded sense is semantics of a completely different sort: “e-semantics”.

The problem is that, as the name suggests, a-semantic accounts are not really semantic at all, and the ubiquity of computability threatens to turn the whole idea of computational neuroscience vacuous. Allowing a neural process to be ”computational” merely on the grounds that it conforms to some algorithm is too weak. What Grush proposes we need are e-semantic accouts - accounts where you can see that:

(2) “The brain (or parts thereof) computes in the sense that it processes information - it deals with what genuinely are information-carrying states – e.g. states that carry information about objects or states of affairs in the environment”. (Grush, 2001, p.158).

So,

> “[I]f there were some principled means to determine which states are representing aspects of the environment […] We would have the means to distinguish those systems that are genuinely computational in the required sense, and there would be no danger of computational neuroscience being assimilated without residue into the general category of computer simulation studies.” (Grush, 2001, p.162).

What we need, then, is a naturalistic theory of how neurons acquire meaning, neurosemantics, and the current brand neurophysiological modeling research isn’t going to provide one (Grush, 2001, pp. 168-169).

The general approach in informational semantics (Dretske, 1981, 1988, Eliasmith 2000, Fodor, 1990, Grush, 1997, 2004, Millikan, 1989, Ryder 2002, 2004, Usher 2001, 2004) is something like this: To get representation out of information you start out with a (physical) account of the dependencies between environmental variables and neural activity. The causally interacting elements (vehicles)[4] of the system to be described as computational (the nervous system) are representations if they are by design isomorphic to something else, and are used in the system to stand in for that something (Cf. Ryder 2002, 2004). Relying on the mathematical notion of isomorphism (see Gallistel 1990 for a worked out example) and perhaps using the statistical notion of mutual information to define “standing in for” (Eliasmith 2000, Usher 2001, 2004) provides means of disambiguating what the individual brain states could stand in for.

This picture has considerable intuitive plausibility, and probably captures much of the logic of representational talk in current neurocience. Yet I will next argue that the

[4] Grush calls them articulants.

account of semantic relations between the brain and the world it gives is incorrect. The problem with this analysis is that it presupposes a very robust sense in which the content of the representation must be "out there", in the environment, and that the very function of neural representations is to reflect with fidelity structures already in the environment (the "veridicality assumption").

## *3. POVERTY OF THE STIMULUS AND THE VERIDICALITY ASSUMPTION*

For the purposes of this paper, I define "poverty of the stimulus" as the hypothesis that the extensions of most concepts do not constitute physical natural kinds, only the concept constitutes a (cognitive) natural kind (e.g. Fodor, 1998). Likewise, by "veridicality assumption" I mean the converse hypothesis that the function of representation is to "pick out" or find out real kinds in the environment, implicitly represented in the information available to the organism via its sensory apparatus[5] (cf. e.g. Millikan 1998, Churchland & Churchland, 2002).

If you are working within the framework of externalist[6] neurosemantics and the semantic notion of computation, then you need to take a stance on the poverty of the stimulus hypothesis. As far as I can tell, *all* current neurosemantics, including Grush's e-semantics, seem inconsistent with the poverty of the stimulus (and instead buy into a very strong veridicality assumption).

[5] This would imply that what are traditionally called secondary qualities are not just illusory, but that their perception is a case of misrepresentation – the system working not as designed. Ryder (2002), for one, explicitly analyses them in these terms.
[6]See footnote 2. All isomorphism based and function-to-stand in for type accounts considered here are externalist in this sense.

The problem with this is that, empirically, it seems plausible that the poverty of the stimulus is true, and should therefore be taken into account as a constraint on philosophical theory construction[7].

Take, for example, color concepts. On the assumption of poverty of the stimulus, you would not consider color properties (categories) such as red, purple and brown as something in the environment, implicit in the structure of the physical energy impinging on an organism. Color is a mind-dependent property[8], only "there" for a particular kind of organism. Assessing the facts about color as a natural phenomenon must therefore take into account not only the physical properties of the environment, but the cognitive properties constituting the point of view of the organism.

From this perspective, in genuinely mental representation there is "more information" in the representation of a domain than is present or available in the domain itself - the distal physical stimuli. This additional information that the "poverty" of the stimulus requires, represents the organism's contribution to the natural phenomenon (say, the ordering relations that define a color space). And insofar as the information constitutes a cognitive (as opposed to physiological) characterization of the organism, it is precisely what sets apart the study of cognitive phenomena and cognitive natural kinds from the study of the physics of the organisms' natural environment and/or brain.

The alternative (which follows from the veridicality assumption) is that you end up with a theory where representation reduces into mirroring or reflecting simple (though possibly quite abstract) *physical* features of the environment, and the organism is charged with finding the true classifications (physically natural ways of organizing the distal environment into objects, categories, properties etc.). Arguably, this isn't at all what brains do – we are capable of representing the environment under non-physical concepts.

---

[7] Of course, it is not possible within the scope of this paper to assess the empirical validity of the poverty of stimulus assumption. Its ubiquity among researchers investigating two forms of cognition that are fairly well understood, namely color perception and natural language, make it at the very least a live philosophical option. I will here simply assess some of its philosophically interesting consequences for neurosemantics.

[8] I.e. a secondary quality.

It is a strong empirical claim that the function of information gathering and processing systems is to find out about physical features of the environment (For a general criticism of “the naturalist camp’s” endorsement of the veridicality assumption, see Akins, 1996).

But this is what you get if you begin with the ”actual referent”, as Grush’s e-semantics and all current neurosemantics seems to do, and then base your *semantics* on causal history, isomorphism and what natural (physical) [9] kind the referents belong to, all that is left for the organism is for it to ”take in” that which is ”given”.

Suppose you had a robust physiological theory that allowed you to identify putative representational vehicles (i.e. neural response classes) in the brain with great accuracy and reliability. To get a semantics assigned to the responses (taking the responses to be representations) you could define the referent as the (physical) stimulus properties to which your vehicles show the greatest statistical dependence with (perhaps by a mutual information measure, as discussed by Eliasmith 2000 and Usher 2001, 2004). In other words you need to generalize to denotation[10]. The veridicality assumption implies that the correct way is to take the denotation of the vehicle to be the physical kind to which the referent belongs.

Millikan (1998, p.57) puts this clearly:

> “It is not a matter of logic, of course, but rather of the makeup of the world, that I can learn from one observation what color Xavier’s eyes are or, say, how the water spider propels itself. It is not a matter of logic that these things will not vary from meeting to meeting. […] most of the knowledge that carries over about ordinary substances is not certain knowledge but merely probable knowledge […] But no knowledge whatever carries over about nonsubstance kinds.”

[9] Taking the distal categories (in the context of neurosemantics) as given under a mentalistic description would be circular, and prejudge precisely the crucial issues; namely, what it is for there to be intensions and what the intension of a particular representation is. The environment must be environment under a physicalistic description.

[10] This is where the disjunction and misrepresentation problems comes in to standard information semantics. The statistical depedendence approach is explicitly designed for this step.

The very representational content of the vehicle would be determined by the real category structure of the world behind the sensory "evidence" available to the organism. Yet, according to the poverty of stimulus hypothesis, there *is* no such *physical* category structure (i.e. there is nothing specific that, say, brown things have in common *physically*). There is no natural class that would allow you to "generalize to denotation" based on a sample of the stimuli belonging to a class (say, brown things). Knowing that a motley selection of objects all fall under the concept BROWN does not enable you to pick out the denotation (all brown things) from the environment without further information, based on just their physical properties (surface reflectances, say)[11].

Not assuming the referents constitute a sample of a real physical kind (poverty of the stimulus) leaves you logically with two options. One would be to assume that they therefore must constitute an "unnatural" (nominal) distal physical category. But this does not seem to offer the right solution. For we are committed to naturalizing cognition, so what we would want is a theory of colors. The other option is to assume that the theory of colors is (in part) a theory of organisms with color vision (standard view in color science), and it seems the natural one.

It does not follow from naturalism, however, that the theory that allows you to generalize to denotation must be a physical one, less so a theory of the distal stimuli. Consider Millikan again (ibid.):

> "There are various contemporary interpretations of the underlying reasons why there are such things as real kinds in nature […] what makes something a natural kind is that there is some such reason: kinds are not natural if they yield inductive knowledge by accident. . […] If a [representation] is to have genuine 'rich inductive potential' it had better attach not just to a pattern of correlated properties, but to a univocal explanatory ground for correlation."

[11] Plus some general striving toward economy of representation.

It is not part of commitment to naturalism – or even externalism - as such, that the source of correlation, or "the univocal explanatory ground for correlation", must be entirely external to the organism's brain. What is required for a kind to be real (for the animal) is for encounters with it to license induction to new encounters. But note that the brain of the organism (and all information within) is an important constant across contexts. The properties need to "carry over" – but what needs to carry over is *represented* properties. Our encounters with a brown object *do* license inductive inferences about further encounters – but this is really as much in virtue of our brain's peculiar color constancy mechanisms, rather than any physical properties of brown objects as such. Indeed, in the case of colors at least, physical properties (e.g. spectral distribution or surface reflectance) are precisely the kinds of things we *cannot* expect to carry over[12].

## 4. CONCLUSIONS

What lessons are we to draw for computational neuroscience and how does this all relate to the distinction between horizontal and vertical semantics, mentioned at the beginning of the paper? Let me take up the distinction between horizontal and vertical semantics first.

One may think of horizontal "information coding" as the causal process of activating an "indicator state" (that is, a neural response class whose activation is causally/statistically dependent on its referent)[13]. Looking at information coding vertically, on the other hand, one looks for entities within the organism which can be considered to represent the value of some variable - the fact that the entity/vehicle was in such-and-such state would "stand

[12] Now, it *could* turn out, as a matter of empirical fact, that the information added to the categorization over and above the sample of brown things and physical regularities in this case would only serve to make the distal categories more *accidental*, or *random*. The additional information needed to specify the appropriate way to generalize to denotation might be a brute incompressible list of reflectances belonging to the category or not. In such a case the contribution of the organism would reflect just so many independent accidents of ontogenetic and phylogenetic history - there would be no *theory* of the contribution of the organism, and color would not turn out to be a natural phenomenon after all. (To be more precise, no theory of colors - say the nature of the cognitive kind BROWN - would be forthcoming, which amounts to the same thing). I see no warrant for such pessimism.

[13] And/or whose structure is designed to be isomorphic with its "intended referent"

for” the fact that the value of the variable was such-and-such - and in genuinely semantic contexts the variables would be elements of a mental representation of some domain. That is, “content (concepts, propositions, properties...*intensions*) coded into the organism”.

Philosophically, these are two quite different ways of looking at representation and information (cf. Cummins, 1996), invoking two very different (and complementary) notions of ”representation”, which perhaps need to be kept separate more carefully than is customary in neuroscience.

Now, a-semantics and c-semantics are both vertical, while Grush’s e-semantics is horizontal. The difference between a-semantics and c-semantics is that in a-semantics the algorithmic computations are a representation of the organism, whereas in c-semantics the organism is a representation of the computations.

In a-semantics, the internal physical configurations of neural tissue and their state transitions can be seen as instantiations of *some* variables and *some* algorithm, and what the algorithm is, is interpreted as it were “bottom up”, starting from the brain states (vehicles). The problem is that, as stated, this “interpretation” does not present the variables and algorithms as part of a mental representation of any domain. Add this requirement, and you get either e-samantics or c-semantics, which (assuming externalism) make essential reference to the environment of the organism, and are therefore genuinely *semantic*.

The difference between c-semantics and e-semantics, on the other hand, is that in e-semantics the organism’s internal states (representations or vehicles), have the function of standing in for something *physical* - something that is really “physically out there” in the environment. The veridicality assumption built into e-semantics requires that the *same* theory (the same formal structure) characterize both the external environment and the organism’s vehicles (isomorphism). c-semantics by contrast makes essential reference to cognitive mechanisms of the organism, that are not merely a reflection of structure in the environment but represent a systematic and genuine contribution of the organism to semantic content. What is represented, content, is a function of *both* the environment and the organism, the contribution of the organism making up for poverty of the stimulus which blocks pure isomorphism/veridicality based accounts.

In sum, if you subscribe to both the poverty of the stimulus and externalist semantics the mental variables represented are not reducible to the physical state of the environment (poverty of the stimulus), but nor do they supervene on the "narrow" state of the individual's brain, either (externalism).

Distinguishing between a-semantics and c-semantics enables us to see that Grush's solution does not exhaust the options. In fact, it could be argued that by confounding a-semantics and c-semantics Grush is blurring precisely that distinction which is crucial to understanding the difference between mere computable regularity and semantic computation - what we set out to elucidate in the first place.

There is, then, an important difference between the "vertical" and the "horizontal" reading of information and that this distinction should probably receive wider recognition than is currently the case in computational neuroscience if it is to have in its foundations a coherent notion of semantic computation. What is more, there are two ways of looking at "going vertical": a-semantics, as discussed by Grush, and c-semantics, of which c-semantics only is a genuinely semantic way of looking at the brain. It can also be maintained that vertical (c) semantics is more "genuinely" semantic than horizontal (e) semantics, at least insofar as e-semantics commits to the veridicality assumption which precludes a systematic "contribution of the organism" to the structure and content of mental representation.

If you accept externalism, then the information (content) encoded in the brain is not a higher level description of the physical processes in the brain. But if you believe in poverty of the stimulus as well, then this content is not a higher level description of the environment, either. Providing an account of what it is would be adequately to face up to the semantic challenge of computational neuroscience. *Pace* Grush, the inadequacy of a-semantics as a semantic theory therefore does not stem from the fact that it is not horizontal ,but that it is vertical in “wrong direction”.

## *REFERENCES*


Akins, K. (1996). Sensory Systems and the Aboutness of Mental States. *Journal of Philosophy*, 93, 337-72.

Churchland, P.S., Koch, C. & Sejnowski, T. (1990). 'What is Computational Neuroscience?' In: Schwarz, E. (ed.): *Computational Neuroscience*. Cambrdge, MA: MIT Press.

Churchland, P.S. & Sejnowski, T. (1992). *The Computational Brain*. Cambridge MA: MIT Press.

Churchland, P.S. & Churchland, P.M. (2002). Neural Worlds and Real Worlds. *Nature Reviews Neuroscience, 3*, 903-907.

Cummins, R. (1989). *Meaning and Mental Representation*. Cambridge MA: MIT Press.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge MA: MIT Press.

Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes*. Cambridge MA: MIT Press.

Eliasmith, C. (2000). *How neurons mean: A neurocomputational theory of representational content*. Ph.D. Thesis. Washington University in St. Louis.

Eliasmith, C. (2003). Moving beyond metaphors: Understanding the mind for what it is. *Journal of Philosophy*, C(10), 493-520.

Fodor, J. (1990). *A Theory of Content*. Cambridge MA: MIT Press.

Fodor, J. (1998). *Concepts*. Oxford: Oxford University Press.

Gallistel, C.R. (1990). *The Organization of Learning*. Cambridge, MA: MIT Press.

Grush, Rick (1995). *Emulation and Cognition*. Doctoral Dissertation, University of California, San Diego.

Grush, R. (1997). The Architecture of Representation. *Philosophical Psychology*, 10, 5-23.

Grush, R. (2001). The Semantic Challenge to Computational Neuroscience. In Machamer, P.K., Grush, R. & McLaughlin, P. (toim.) *Theory and Method in the Neurosciences* (pp. 155-172). Pittsburgh, University of Pittsburgh Press.

Grush, R. (2004). The Emulation Theory of Representation: Motor Control, Imagery and Perception. *Behavioral and Brain Sciences*, 27, 377-442.

Koch, C. (1990). Biophysics of Computation: Toward the Mechanisms of Information Processing in Single Neurons. In: Schwarz, E. (ed.): *Computational Neuroscience*. Cambrdge, MA: MIT Press.

Millikan, R.G. (1989). Biosemantics. *Journal of Philosophy*, 86, 288-302.

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*, 21, 55-100.

Piccinini, G. (2004a). Functionalism computationalism and mental contents. *Canadian Journal of Philosophy* 34: 375-410.

Piccinini, G. (2004b). "Functionalism, Computationalism, and Mental States," *Studies in the History and Philosophy of Science*.

Putnam, H. (1975). The Meaning of "Meaning". In Gunderson, K. (Ed.) Language, Mind and Knowledge. *Minnesota Studies in the Philosophy of Science*, vol. 7. (pp. 131-193) Minneapolis: University of Minnesota Press, 1975.

Ryder, D. (2002). *Neurosemantics: A Theory*. University of North Carolina at Chapel Hill.

Ryder, D. (2004). SINBAD Neurosemantics: A Theory of Mental Representation. *Mind & Language*, 19, 211-241.

Sejnowski, T., Koch, C. & Churchland, P.S. (1988) 'Computational Neuroscience' *Science*, 241.

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

Usher M. (2004). Comment on Ryder's SINBAD Neurosemantics: Is Teleofunction Isomorphism the Way to Understand Representations? *Mind and Language*, 19, 241-248.