O'Brien, Gerard Joseph; Opie, Jonathan Philip.
Vehicle, process, and hybrid theories of consciousness, *Behavioral and Brain Sciences*, 2004; 27 (2):303-305.

tion (3) we must also drop the assumption of identity between explicitness and consciousness. As several commentators suggested (Church, Cleeremans & Jimenez, Dennett & Westbury, Kurthen, McDermott, Van Gulick, and Wolters & Phaf; cf. *BBS* 22[1], 1999), explicitness could be necessary but not sufficient for consciousness. O&O's only support for the identification of explicitness and consciousness comes from their reappraisal of the dissociation studies. However, this is possibly the most questionable point in their paper. Lacking a final verdict on the issue, it seems that their persistence in identifying both properties is due to their thinking that "it is clearly incompatible with the connectionist vehicle theory of phenomenal experience [to assume] the operation of explicitly represented information that does not figure in consciousness" (p. 187). I claim that there is no such incompatibility, insofar as we drop structural explicitness. This leads us to position (4).

First, all that a vehicle theory of consciousness demands, according to Thomas & Atkinson and Van Gulick (cf. *BBS* 22[1], 1999), is a principled distinction between kinds of representations, R and R^1, so that the intrinsic properties of a given kind make it the basis of conscious experience. Second, from a process explicitness viewpoint, the more accessible some information I is, the more explicit I will be. Third, from a vehicle consciousness perspective we can say that I becomes conscious only when it is explicit and encoded by a specific kind of representation, (say, R^1). This would fill position (4).

A connectionist version of this possibility is: (1a) Two kinds of patterns, P and P^1. (2a) Gradation of explicitness: information in weights is less accessible than information in patterns. (3a) Information in patterns is not immediately conscious; only some patterns are so, say P^1. But now we open the door to classical vehicle theories of consciousness: (1b) Two kinds of symbols, S and S^1. (2b) Gradation of explicitness: some symbolic information is more explicit by being more accessible. (3b) Only an explicit S^1 makes its contents conscious.

Both versions, however, face the same problem: how to single out an intrinsic property that provides a principled distinction between the patterns P and P^1 or the symbols S and S^1. There is an obvious place to look for such a principled structural distinction between representational kinds: the distinction itself between patterns and symbols. Suppose that we allow both kinds of representations in our system. We can fill position (4) as follows: (1c) Two kinds of representations: symbols and patterns. (2c) Gradation of explicitness: from content in weights to content in patterns, and content in accessible symbols. (3c) Content is conscious only when it is rendered into explicit symbolic format. This can require the extraction of the content from the network.

Two notes: First, a "purely vehicle" theory of consciousness need not be "purely connectionist" or "purely symbolic"; it can contain instances of both representational kinds. Second, even if the content has to be extracted for being conscious, this does not make it a process theory. It is not being extracted that makes the content conscious; it is being symbolic that makes it so. If O&O insist that extraction makes this version a process theory of consciousness, then they should equally answer the charge (Mac Aogáin, Wolters & Phaf; cf. *BBS* 22[1], 1999) that a pattern is always the product of some process.

Things are probably much more mixed up than suggested by any simple theory of consciousness. If connectionist and symbolic vehicles belong to different "representational genera" according to the contents they are capable of representing (Haugeland 1991), then they may underlie different kinds of conscious states. On the other hand, it is also dubious that a purely vehicle or a purely process theory will account for consciousness. I have argued elsewhere (Martinez & Ezquerro 1998) that intuitions from the structural and the process views should be integrated to offer an appropriate characterization of explicitness, and an analogous claim can be made with respect to vehicle and process theories of consciousness. In other words, the character of conscious experiences may depend not on what a representation is or on what it does but rather in the subtle interaction of both factors.

ACKNOWLEDGMENT

Preparation of this paper was supported by the MCyT research project BFF2002–03842.

Authors' Response

Vehicle, process, and hybrid theories of consciousness

Gerard O'Brien and Jonathan Opie

Department of Philosophy, School of Humanities, University of Adelaide, South Australia 5005, Australia. **gerard.obrien@adelaide.edu.au**
jon.opie@adelaide.edu.au
http://www.arts.adelaide.edu.au/humanities/gobrien/
http://www.arts.adelaide.edu.au/humanities/jopie/

Abstract: Martínez-Manrique contends that we overlook a possible nonconnectionist vehicle theory of consciousness. We argue that the position he develops is better understood as a hybrid vehicle/process theory. We assess this theory and in doing so clarify the commitments of both vehicle and process theories of consciousness.

In developing the connectionist vehicle theory of phenomenal experience we were mindful of two things: (1) that consciousness is, by and large, a consequence of the brain's representing activity, (2) that current theories of mental representation are heavily influenced by the classical computational theory of mind. Connectionism presents a unique opportunity to rethink consciousness because, unlike classicism, its account of cognition is framed in terms of certain structural properties of the brain. In particular, connectionism distinguishes between two structurally distinct kinds of representing vehicle: connection weight representations, and activation pattern representations. Others have noticed the possibility of identifying phenomenal experience with the relatively transient activation patterns that constantly course across the brain, while assigning connection weights the twin tasks of information storage and computational substrate (Rumelhart et al. 1986, p. 39; Smolensky 1988, p. 13; Lloyd 1991; 1995; 1996). In our target article we sought to further develop and defend this idea, conjecturing that phenomenal consciousness is identical to the vehicles of explicit representation in the brain – such vehicles being understood as stable patterns of neural activation.

Martínez-Manrique, in his useful commentary, argues that we have overlooked a possible variety of vehicle theory, one moreover that contains both connectionist and classical elements. His crucial move, in canvassing this possibility, is to exploit the distinction between structural and process conceptions of explicit representation. In our target article we develop a generic representational framework that characterizes explicit representation in structural terms. Martínez-Manrique observes that there is well-known analysis, primarily due to Kirsh (1990), according to which information is explicit if it is readily accessible by a cognitive system, and is, by degrees, less explicit if it is more difficult to access. As Martínez-Manrique admits, this is a process conception of explicit representation. But one may

recover a vehicle theory of consciousness, he thinks, if explicitness is treated as necessary but not sufficient for consciousness. An additional (vehicle) criterion might be added, to the effect that a widely available representational content will be conscious when its vehicle satisfies some intrinsic, structural constraint. This, claims Martínez-Manrique, ultimately permits a vehicle theory in which connectionist (activation pattern) and classical (symbolic) representations both play a part.

At the outset we must say that **Martínez-Manrique**'s analysis of the space of possible theories seems to us seriously flawed. Contrary to what he claims, one *cannot* coherently combine a vehicle theory of consciousness with a process conception of explicit representation. A vehicle theory of consciousness seeks to explain phenomenal experience in terms of the *intrinsic* nature of the brain's explicit representing vehicles – in terms of what these vehicles *are* rather than what they *do.* A process conception of explicitness holds that information is explicitly represented in a cognitive system when it can be easily accessed. But the ease with which a representational content can be accessed is not solely or even largely determined by the intrinsic properties of the vehicle that carries it; it is determined by the nature of the cognitive system in which that vehicle is embedded. Consequently, there just is no coherent formulation of a vehicle theory of consciousness which adopts a process conception of explicitness: one cannot hope to explain phenomenal consciousness in terms of intrinsic properties of the brain's explicit representing vehicles when explicitness is determined largely by properties extrinsic to these vehicles. We thus hold to our conclusion, drawn in our target article, that only connectionism has the resources to develop a plausible vehicle theory of consciousness.

Given this, perhaps a better interpretation of **Martínez-Manrique**'s commentary is not that there is a nonconnectionist vehicle theory we have overlooked but that there is a way of combining structural and process criteria within a single account – a maneuver which, in effect, generates a hybrid vehicle/process theory. Martínez-Manrique's ultimate suggestion is that a content is conscious "only when it is rendered into explicit symbolic format" (para.8). Being symbolic is the vehicle criterion. What is the process criterion? In typical process accounts a representational content is taken to be conscious when its vehicle is subject to relations of widespread informational access – that is, when it has rich and widespread information processing *effects* on the brain's ongoing operations. However, as we explained previously (O'Brien & Opie 1999, pp. 176–77), any hybrid account that followed this line would violate one of the deepest intuitions we have about consciousness: that *conscious experience makes a difference.* If a symbolic content must give rise to widespread information processing effects in order to enter consciousness, its being conscious cannot be the cause of those effects. But this is not Martínez-Manrique's strategy. Rather than focusing on informational *access,* his process criterion is informational *accessibility:* representational contents are conscious when they are encoded symbolically and can readily be accessed and put to use in the service of cognition. And it might be argued that this change of focus renders his hybrid vehicle/process theory consistent with the causal potency of consciousness.

One obvious problem with any theory that makes informational accessibility, rather than informational access, criterial for consciousness is that it runs the risk of being empirically implausible. Nothing could be clearer than the fact that we have at our fingertips a vast store of unconscious but readily accessible information. **Martínez-Manrique**'s proposal can skirt over this difficulty, however, because it holds that accessibility is insufficient for consciousness; consciousness also requires the satisfaction of a structural (vehicle) criterion. Our worry with this hybrid vehicle/process theory is different, but just as straightforward. We think it unmotivated and unparsimonious. It is unmotivated because, although it is clear why one might seek to explain consciousness by identifying it with either the intrinsic properties of the brain's representing vehicles (in doing so one connects consciousness with the very entities that drive human cognition) or the information processing effects of these representing vehicles (in doing so one connects consciousness with the process of accessing the information these vehicles carry), it is unclear why one would seek to explain consciousness in terms of the fact that certain representational contents are more readily accessible than others. And it is unparsimonious because it accounts for consciousness in terms of both intrinsic and extrinsic properties of the brain's representing vehicles when simpler theories that restrict themselves to one or other class of properties have yet to be fully explored.

In this vein, it is useful to consider why **Martínez-Manrique** so quickly dismisses our connectionist vehicle theory. He does so because he thinks connectionism is incapable of distinguishing conscious representing vehicles from their unconscious counterparts by recourse to a structural criterion of explicitness. And Martínez-Manrique reaches this conclusion by interpreting the *stability* of an activation pattern representation as a *temporal,* rather than a *structural,* property of a neural network. We think Martínez-Manrique is wrong about this. As we were at pains to point out in our original "Authors' Response" (O'Brien & Opie 1999r, pp. 181), there is a widespread misunderstanding of the significance of stability in connectionist networks that issues from a failure to distinguish between the behavior of real neural networks and the properties of their digital simulations. Since this error persists, we will conclude our discussion by briefly revisiting this issue.

In a simulation, a neural network's activity is modeled as an array of *numerical activation values,* which are periodically updated by algorithms that model the network's internal processes. Simulated relaxation search thus proceeds via a sequence of determinate numerical arrays, giving the impression that prior to stabilization a neural network jumps between specific points in its activation space, and hence generates a sequence of short-lived activation patterns before settling into a longer lasting pattern. This is the picture **Martínez-Manrique** has in mind when he claims that there is no intrinsic structural distinction among the "transient" patterns that precede the production of a "stable" pattern, and hence no structural criterion which can ground a distinction between unconscious and conscious states (para. 3). But this picture is misleading. Whenever one employs a numerical value to describe a continuously variable physical property, one is imposing an instantaneous value on this property. Since neural spikes are discrete events, neural spiking rates *do not have instantaneous values;* the notion of a rate, in this case, only makes sense relative to some time window. In a real network, stabilization is a process in which constituent neurons adjust the absolute timing of their spikes until a determinate firing rate

is achieved. Prior to stabilization, neural networks do not jump around between points in activation space. Stabilization is the process whereby a network first generates a determinate activation pattern, and thereby *arrives* at a point in activation space.

So a real neural network does not generate a pattern of activation, and thus a determinate representational content, until it achieves some measure of stability. Consequently, there is no distinction between "stable" and "transient" activation patterns. Stable activation patterns are physical objects, objects moreover that are structurally distinct from a neural network's configuration of connection weights. And it is this distinction, between activation pattern representation and connection weight representation, that according to our vehicle theory marks the boundary between the conscious and the unconscious.

References

[The letter "r" before author's initials indicates Response article references]

Haugeland, J. (1991) Representational genera. In: *Philosophy and connectionist theory,* ed. W. Ramsey, S. P. Stich & D. E. Rumelhart. Erlbaum. [FM-M]

Kirsh, D. (1990) When is information explicitly represented? In: *Information, language and cognition,* ed. P. Hanson. University of British Columbia Press. [FM-M, rGO]

Lloyd, D. (1991) Leaping to conclusions: Connectionism, consciousness, and the computational mind. In: *Connectionism and the philosophy of mind,* ed. T. Horgan & J. Tienson. Kluwer. [rGO]

(1995) Consciousness: A connectionist manifesto. *Minds and Machines* 5:161–85. [rGO]

(1996) Consciousness, connectionism, and cognitive neuroscience: A meeting of the minds. *Philosophical Psychology* 9:61–79. [rGO]

Martínez, F. & Ezquerro, J. (1998) Explicitness with psychological ground. *Minds and Machines* 8:353–74. [FM-M]

O'Brien, G. & Opie, J. (1999) A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences* 22(1):127–96. [FM-M, rGO]

O'Brien, G. & Opie, J. (1999r) Putting content into a vehicle theory of consciousness. (Author's Response to Open Peer Commentary.) *Behavioral and Brain Sciences* 22(1):175–96. [rGO]

Rumelhart, D. E., Smolensky, P., McClelland, J. L. & Hinton, G. E. (1986) Schemata and sequential thought processes in PDP models. In: *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2: Psychological and Biological Models,* ed. J. L. McClelland & E. E. Rumelhart. MIT Press. [rGO]

Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11:1–23. [rGO]

Commentary on **Anne Campbell (1999). Staying alive: Evolution, culture, and women's intrasexual aggression. BBS 22(2):203–252.**

Abstract of the original article: Females' tendency to place a high value on protecting their own lives enhanced their reproductive success in the environment of evolutionary adaptation because infant survival depended more upon maternal than on paternal care and defence. The evolved mechanism by which the costs of aggression (and other forms of risk taking) are weighted more heavily for females may be a lower threshold for fear in situations which pose a direct threat of bodily injury. Females' concern with personal survival also has implications for sex differences in dominance hierarchies because the risks associated with hierarchy formation in non-bonded exogamous females are not off-set by increased reproductive success. Hence among females, disputes do not carry implications for status with them as they do among males, but are chiefly connected with the acquisition and defence of scarce resources. Consequently, female competition is more likely to take the form of indirect aggression or low-level direct combat than among males. Under patriarchy, men have held the power to propagate images and attributions which are favourable to the continuance of their control. Women's aggression has been viewed as a gender-incongruent aberration or dismissed as evidence of irrationality. These cultural interpretations have "enhanced" evolutionarily based sex differences by a process of imposition which stigmatises the expression of aggression by females and causes women to offer exculpatory (rather than justificatory) accounts of their own aggression.

Hierarchy disruption: Women and men

János M. Réthelyi and Mária S. Kopp

Institute of Behavioral Sciences, Semmelweis University, Budapest, 1089, Hungary. **retjan@net.sote.hu kopmar@net.sote.hu**
www.behsci.sote.hu

Abstract: The application of evolutionary perspectives to analyzing sex differences in aggressive behavior and dominance hierarchies has been found useful in multiple areas. We draw attention to the parallel of gender differences in the worsening health status of restructuring societies. Drastic socio-economic changes are interpreted as examples of hierarchy disruption, having differential psychological and behavioral impact on women and men, and leading to different changes in health status.

Campbell's (1999) target article about gender differences in aggression and status-seeking behavior describes a convincing body of evidence and presents a plausible evolutionary explanation. The target article and the commentaries raise a number of questions concerning the consequences and practical implementations of an evolutionary theory. We propose that several new findings in the areas of epidemiology and health psychology yield parallel results that fit well with Campbell's model. The phenomenon of health status deterioration in restructuring societies, primarily those of Central and Eastern Europe, and the until-now not convincingly explained gender differences in health deterioration are results that could serve as a bridge between a behaviorally oriented evolutionary model and large-scale epidemiological findings. Reading the article and the following debate was a profound intellectual experience; the recognition of parallel results between different fields was even more exciting.

Socio-economic changes following political transition in the countries of Central and Eastern Europe have influenced people's lives in a variety of ways. Among these phenomena, one of the most striking is the declining health status of these societies (Feachem 1994). The dynamics of the process show different characteristics in different countries according to the chronological nature of the political changes. In Hungary, deterioration began in the early 1970s at a constant slow grade, and male life expectancy decreased by three years between 1970 and 1995, parallel with political softening and the beginning of economic polarization (Bobak & Marmot 1996; Kopp 2000). As a more severe