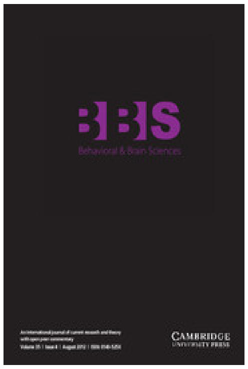

Theory of mind and other domain-specific hypotheses

C. M. Heyes

Behavioral and Brain Sciences / Volume 24 / Issue 06 / December 2001, pp 1143 - 1145
DOI: 10.1017/S0140525X01270149, Published online: 17 December 2002

Link to this article: http://journals.cambridge.org/abstract_S0140525X01270149

How to cite this article:
C. M. Heyes (2001). Theory of mind and other domain-specific hypotheses. Behavioral and Brain Sciences,24, pp 1143-1145 doi:10.1017/S0140525X01270149

BEHAVIORAL AND BRAIN SCIENCES (2001) **24,** 1135–1148
Printed in the United States of America

Continuing Commentary

Commentary on **Luiz Pessoa, Evan Thompson, and Alva Noë (1998) Finding out about filling-in: A guide to perceptual completion for visual science and the philosophy of perception. BBS 21:723–802.**

Abstract of the original article: In visual science the term *filling-in* is used in different ways, which often leads to confusion. This target article presents a taxonomy of perceptual completion phenomena to organize and clarify theoretical and empirical discussion. Examples of boundary completion (illusory contours) and featural completion (color, brightness, motion, texture, and depth) are examined, and single-cell studies relevant to filling-in are reviewed and assessed. Filling-in issues must be understood in relation to theoretical issues about neural–perceptual isomorphism and linking propositions. Six main conclusions are drawn: (1) visual filling-in comprises a multitude of different perceptual completion phenomena; (2) certain forms of visual completion seem to involve spatially propagating neural activity (neural filling-in) and so, contrary to Dennett's (1991; 1992) recent discussion of filling-in, cannot be described as results of the brain's "ignoring an absence" or "jumping to a conclusion"; (3) in certain cases perceptual completion seems to have measurable effects that depend on neural signals representing a presence rather than ignoring an absence; (4) neural filling-in does not imply either "analytic isomorphism" or "Cartesian materialism," and thus the notion of the bridge locus – a particular neural stage that forms the immediate substrate of perceptual experience – is problematic and should be abandoned; (5) to reject the representational conception of vision in favor of an "enactive" or "animate" conception reduces the importance of filling-in as a theoretical category in the explanation of vision; and (6) the evaluation of perceptual content should not be determined by "subpersonal" considerations about internal processing, but rather by considerations about the task of vision at the level of the animal or person interacting with the world.

Visuo-cognitive disambiguation of occluded shapes

Rob van Lier
Nijmegen Institute for Cognition and Information (NICI), University of Nijmegen, The Netherlands. **r.vanlier@nici.kun.nl**
www.nici.kun.nl/People/LiervanRJ/index.html/

Abstract: Pessoa et al. (1998a) underexposed the broad and rich variety of stimuli in the amodal completion domain. The disambiguation of occluded shapes depends on very specific figural properties. Elaborations on such disambiguations of rich and complex stimuli, tied up with a visuo-cognitive origin of amodal completion, further position Pessoa et al.'s considerations on neural filling-in and the personal-subpersonal distinction.

Pessoa et al. (1998a) have written an impressive paper on perceptual completion, comprising a wide scope of phenomena with diverse phenomenological qualities. Amodal completion certainly belongs to the weakest of the discussed completion phenomena. Every observer can witness that the phenomenological presence of amodal contour completions is not as compelling as, for example, blind spot filling-in, neon color spreading, or even the perception of illusory contours. Nevertheless, convincing psychophysical data exist on the relevance of amodal contour completions as well. Much of the research in the domain is concerned with the disambiguation of occluded shapes. That is, while in fact an infinitive number of different completions are possible for each and every visual pattern, only a few of them are plausible. Although Pessoa et al. briefly mentioned that local and global figural properties may influence completion, the richness of the domain, the diversity of completions, and the vision-versus-cognition dilemma (typical of this domain), are underexposed and not related to their own concepts (e.g., the personal-subpersonal distinction). This is a missed opportunity.

Whereas in local approaches, completion depends on specific local configurations and proceeds by way of curved or linear interpolation between contour ends, in global approaches specific overall shape regularities (like bilateral symmetries) determine completion.[1] It is important to note here that local and global strategies may converge to the same shape but may also diverge to different shapes. The relevance of and competition between, both types of completions have been investigated by, for example, Sekuler (1994), Sekuler et al. (1994), and Van Lier et al. (1994; 1995a; 1995b).[2] The stimulus-dependent plausibility of a small set of completions is not as self-evident as it might seem to be. For example, with his "ignorance-of-absence" assumption Dennett (1991) also disregards the specific influence of figural properties on completion by stating that the brain jumps to a conclusion (the issue here would be: what conclusion?).

The unique status of a small set of completions also holds if the stimulus domain is further extended, for example, to a completion of the unseen back of a nonfamiliar object (Van Lier & Wagemens 1999) or so-called fuzzy completions of quasi-irregular shapes (Van Lier 1999). The more enriched the stimuli (see also Fig.1), the more compelling the question of whether we are (still) dealing with *visual* completions. Would there be a fundamental difference between the visual and/or cognitive processing of, say, a partly occluded square and the back of a tree trunk (Van Lier 1999)? Is a simple "vision/cognition" verdict possible anyway?

As briefly indicated by Pessoa et al., the vision-versus-cognition issue has been part of the debate in the literature. In particular, global completions are often thought to be cognitive (e.g., Kanizsa 1985). So far, however, psychophysical evidence for the relevance of local and global amodal completions does not rigorously separate visual processing from cognitive processing. In addition, investigating this issue in terms of neural activation in the visual cortical area will turn out to be a hazardous enterprise – not only because the presumed mapping between the responsible neural substrate and the specific completion is unclear (see also Pessoa et al.), but also because there is simply no clear-cut border where the visual system ends and the cognitive system starts.[3] So, even if there were evidence on the neural filling-in of the

0140-525X/01 $12.50

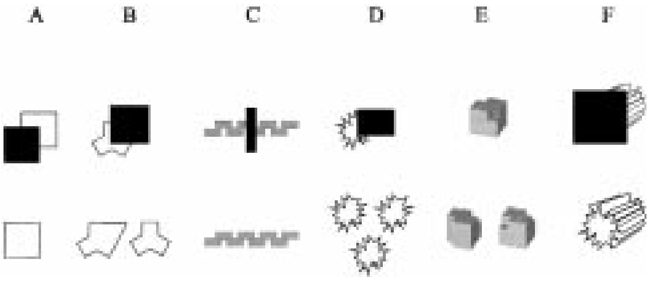

Figure 1 (van Lier). **A:** Global-local convergence. **B:** Global-local divergence. **C:** Global completion in the absence of a local alternative. **D:** Fuzzy completion with many plausible completions. **E:** Two plausible completions of the back of an object (after 90° rotation about the vertical). **F:** Fuzzy object completion. Although only very little is visible, most observers have a clear idea about the object's shape.

unseen back of a tree trunk, it would not necessarily resolve the vision-cognition dilemma.

What really needs careful investigation in connection with amodal completion, is the convergence from an infinitive number of possible completions to a small class of preferred completions, the intra/interobserver consistency on these completions, and their relation to relevant figural properties. The necessity of an entangles visuo-cognitive account of completions, being an inextricable part of our interpretation of the surrounding world, comes close to Pessoa et al.'s personal-level account. However, the apparent systematic relation between preferred completions and particular figural aspects calls for investigation into the underlying visuo-cognitive mechanisms which inevitably touch the subpersonal level Pessoa et al. find less important.

ACKNOWLEDGMENT
RVL was supported by a NOW/MAG PPS grant and a KNAW grant.

NOTES

[1] Pessoa et al. incorrectly classified the Wouterlood and Boselie's (1992) "local" approach as "hybrid" (which would account for global and local aspects).

[2] Within the context of Structural Information Theory, the approach of Van Lier et al. (1994) incorporates both global and local aspects in contrast with the earlier approach of Buffart et al. (1981).

[3] Top-down activations in the visual cortical area as in, for example, mental imagery, or (still to be unraveled) complex cortical "loops," would only further diffuse such a distinction.

Perceptual filling-in and the resonant binding of distributed cortical representations

Tony Vladusich
Centre for Visual Science, Research School of Biological Sciences, Canberra, ACT, 2601, Australia. **vladusich@rsbs.anu.edu.au**

Abstract: Pessoa et al. (1998a) summarize a wide body of data suggesting that perceptual filling-in phenomena can be attributed to neural filling-in processes. However, they reject, on philosophical grounds, the hypothesis that filled-in representations in the brain are the immediate substrate of visual percepts. It is proposed in this commentary that resonant binding between distributed cortical areas may instead be the crucial ingredient for conscious visual percepts, and that filling-in processes may facilitate the interactions between behaving organisms and object surfaces. These suggestions circumvent some of the philosophical problems associated with the idea of localized visual representations.

The recent target article of Pessoa et al. (1998) on the relationship between perceptual filling-in and its putative neural substrates has provoked a lively debate (see accompanying commentaries and response). Perceptual "filling-in" is the attribution of visual properties, such as form, color, texture, or motion, to regions of the visual field (e.g., retinal blind spot) that receive little or no direct sensory stimulation (Ramachandran 1992). Pessoa et al. outline a wide body of evidence favoring the conclusion that perceptual filling-in can be identified with a spatial (topographic) spreading of neural activity in the brain (i.e., neural filling-in). However, the authors reject the idea of a one-to-one "isomorphic" relationship between perception and neural activity; that is, "an ultimate neural foundation in which an isomorphism obtains between neural activity and the subject's experience" (p. 742). One reason for rejecting the idea of isomorphism is that it implies that there is a "final stage" in the brain where neural activity maps directly onto perceptual states; "Why must there be one particular neural stage whose activity forms the immediate substrate of visual perception?" (p. 742). This objection is certainly not new, having featured prominently in Gibson's (1979) arguments against the notion of perception as a form of representation. Indeed, Gibson's direct theory of visual perception, and his notion of reciprocal interactions between perceiving organisms and environment, were derived as alternatives to the notion of neural representation (see Ullman 1980).

A related problem with the idea of isomorphic neural filling-in concerns the impression that the brain is somehow "painting" internal pictures of the external world. Inevitably the question arises as to who or what is "looking at" the internal panoramic canvas. This type of reasoning suffers from the paradox that neural mechanisms in the brain are attributed to the qualities of entire (perceiving) organisms (i.e., the qualities of perception). Gibson (1979) circumvented this paradox by suggesting that the proper subject of perception is the whole organism interacting with its environment, not its constituent representational apparatus. Such an idea has the advantage of "closing the loop" between the perceiver and what is perceived.

Pessoa et al. themselves adopt what might be termed a "neo-Gibsonian" perspective of filling-in (cf. Marr 1982); "the task of vision is not to produce representations from images, but rather to discover through the perceptual system what is present in the world and where it is" (p. 744). The authors claim that "neural filling-in helps the animal find out about its environment" (p. 790) by integrating local neural responses at object edges into global estimates of surface properties (form, color, texture, and motion). To avoid the pitfalls of isomorphism, Pessoa et al. suggest that neural filling-in "is sufficient to *produce* the percept, but not to *constitute* it" (p. 787), meaning that perception "is inherently world-involving and therefore cannot be reductively identified with neural states inside the individual" (p. 788). This suggestion might profitably be viewed as a neo-Gibsonian theory wherein perception involves the moment-to-moment interactions or "resonances" between animal and environment. However, the idea that perception is ultimately "world-involving" has traditionally been a difficult one to assess (Ullman 1980), possibly because notions like animal-environment resonances offer little in the way of putative physical mechanisms (cf. Grossberg 1980).

This commentary advocates a framework which might serve to alleviate the respective difficulties associated with (1) the putative relationship between filling-in and visual perception, and (2) Gibson's insights concerning reciprocal animal-environment interactions. In particular, we consider how the outputs of a process like neural filling-in might facilitate the effective control of behavior in a "closed loop" dynamical system. The question is: How might functionally specialized brain regions work together as a single functional unit to control behavior? One plausible candidate is the dynamic "binding" of distributed cortical events through reciprocal exchange of information, or resonance (e.g., Grossberg 1995). As Pessoa et al. point out, "brain regions are not independent stages or modules, but rather interact reciprocally . . . the brain relies on distributed networks that transiently coordinate their activities" (p. 742). The implication here is that *visual perception* is not a localized product of any single brain region or neural representation, but rather emerges from the dynamic interactions between distributed brain regions. Visual perception may therefore be viewed as an *emergent property* of reciprocal information flow between multiple brain regions. While this idea is certainly not new (Grossberg 1980), a few points can be made to clarify and extend the position.

Since the resonant binding of distributed codes in the brain requires reciprocal anatomical connections between functional areas involved in (say) visual processing and motor control, the brain as a whole can "close the information loop" in a unitary action-perception cycle. This closed loop system effectively exorcises the "homunculus" from the perceptual system, and confers no privileged status on any particular neural representation, filled-in or otherwise. Consistent with the suggestions of Pessoa et al., filling-in or spatial integration may serve an important role in the guidance of behavior by enabling an organism to interact with entire object surfaces, without admitting any special significance to the filling-in process itself. Further, the current view clarifies how brain regions involved in motor control and three-dimensional (3D) spatial representation can guide and modify perception by means of "attentive looking" (Whittle 1998), involving feedback (e.g., outflow commands) to occipital brain regions involved in "bottom-up" visual processing. Spatial attention is thereby allocated to objects that are relevant to the behavior of an organism. This proposal also suggests how object "affordances" (Gibson 1979), the potential behavioral actions afforded by objects, might arise by means of resonant binding between outflow motor commands in frontal cortex and spatial/object representations in occipital, parietal, and temporal cortex.

The proposal outlined above is consistent with data from lesion studies. For example, normal vision requires an intact temporal cortex, as evidenced by the difficulty individuals suffering from damage in the temporal lobe sometimes have in binding together distributed object features into coherent objects. As noted in the commentary of Walker and Mattingley (1998), patients with parietal lesions fail to perceive, or neglect, regions of the visual field contralateral to the cortical lesion. These patients, in effect, lose their capacity for "attentive looking" (as distinct to physical looking). Such evidence clearly indicates that visual perception requires the functional integrity of distributed representations that are not confined to the occipital lobe. Another example of the significance of distributed representations in visual perception comes from brain-imaging studies demonstrating that visual phenomena, such as the McCollough effect (McCollough 1965), are associated with activity in widely distributed brain regions, some of which (e.g., prefrontal cortex) are not classically associated with visual perception (e.g., Barnes et al. 1999). The illusory colors in the McCollough effect are known to undergo filling-in (see Broerse et al. 1999), leading to the suggestion that interactions between filling-in representations and "cognitive" (prefrontal) representations may interact during the illusion. From this perspective, it is plausible to suggest that measurable effects on visual perception could be generated by (say) lesions to the prefrontal cortex, perhaps involving a disruption in the capacity for "attentive looking" or visually guided exploration.

Authors' Response

Filling-in: One or many?

Luiz Pessoa,[a] Evan Thompson,[b] and Alva Noë[c]

[a]*Department of Computer and Systems Engineering, Center of Technology, Federal University of Rio de Janeiro, Ilha do Fundao, Rio de Janeiro, RJ 21945–970, Brazil;* [b]*Department of Philosophy and Centre for Vision Research, York University, North York, Ontario, Canada M3J 1P3;* [c]Department of Philosophy, University of California, Santa Cruz, Santa Cruz, CA 95064. **pessoa@cos.ufrj.br** **www.cos.ufrj.br/~pessoa/** **event@yorku.ca** **www.yorku.ca/research/vision/evant.html/** **anoe@cats.ucsc.edu** **www2.ucsc.edu/people/anoe/**

Abstract: (1) The main issue with regard to modal and amodal completion is not which phenomena are cognitive, and which perceptual. At the level of the animal, both are visuo-cognitive. At the level of visual processing, however, we need to dissect the different functional effects of these kinds of completion. (2) Resonant binding between distributed cortical areas may play a role in perceptual completion, but evidence is needed.

Consider two extreme perceptual completion situations: (1) a Kanizsa triangle, in which the illusory triangle appears with marked contours and as visibly brighter than the white area around it (which has the same luminance); and (2) an object, such as a cat, appearing behind a picket fence. In the first case, we see the triangle and the contours. In the second, we perceive a cat (not merely the head and rear). In the first case, there is good evidence that mechanisms of contour integration at early visual stages play an important role in the perception of the illusory contours (Peterhans & von der Heydt 1989; von der Heydt & Peterhans 1989). In the second, although early mechanisms undoubtedly play an important role, most likely a host of other processes are involved. Granted that the attempt to divide the mechanisms involved in these two cases into distinct "perceptual" and "cognitive" categories is unproductive, should the two cases be treated equivalently?

Van Lier argues that because there is no clear-cut border where the visual system ends and the cognitive system begins, we should think more generally of visuo-cognitive completions or behaviors. Our answer is yes, and no.

First, the yes part. As we argue in our target article (Pessoa et al. 1998a), the point of vision at the level of the animal or subject is to bear witness to what goes on in the world, not to gauge what goes in the head when one perceives. A crucial feature of ordinary perceptual experience is its transparency: perception aims directly at the world and does not ordinarily involve beliefs about what goes on in the visual system (Pessoa et al. 1998a; Noë et al., in press; Noë & Thompson, in press). According to the enactive approach to perception, what Marr (1982) called the computational task of vision, is not the production of internal world-models, but rather the guidance of action and exploration (Noë et al., in press; Pessoa et al. 1988a; Thompson et al. 1992; Varela et al. 1991; see also Ballard 1991; Clark 1996). The subject of vision, in this way of thinking, is not the early-vision, information-processing stream, but rather the whole, environmentally-situated animal, actively engaged in movement and exploration. At the level of the whole animal, perception, cognition, and action are interdependent capacities. Therefore, to separate these capacities seems futile, not only in the case of perceptual completion, but in general. For this reason, when considering perceptual completion at the level of the animal or person, we completely agree with **van Lier** on the "necessity of an entangled visuo-cognitive account of completions." Indeed, this point reinforces our point in the target article that perceptual content needs to be understood at the level of the animal or person acting in the world, for it implies that we need to think of the visual system not as an encapsulated and "cognitively impenetrable" subpersonal module (Pylyshyn, in press), but rather as a system that is fully integrated into the life of the cognitively endowed animal (see Noë et al., in press).

But how should we study perceptual completion? Here we disagree with **van Lier**. As we argue in the target article (Pessoa et al. 1998a) and our "Authors' Response" (Pessoa et al. 1998b), perceptual completion is not one phenomenon, but many under the same heading. Without a careful separation of these phenomena, we risk lumping together a number of different things. Is the brain simply "ignoring an absence," as some researchers argue for the blind spot (Dennett 1991; Kranda 1998; Neumann 1998)? Or is it engaging topographically organized, early visual circuits that mitigate the local indeterminacy of early measurements through short-range interaction (see Pessoa et al. 1998a; Pessoa & Neumann 1998)? In the context of amodal completion (see Rensink & Enns 1998), does the completion process itself posit new visual elements (e.g., extending contours and filling in surfaces)? Or does it simply impose a nonvisual structure onto elements already present? In this context, we argue that a fruitful strategy is to investigate the functional effects of completion. For example, if illusory contours rely on the same kinds of mechanisms as real contours, then we should expect to observe that aftereffects transfer between them. This is exactly what is observed (Berkeley et al. 1994; Paradiso et al. 1989; see also sects. 6.12 and 7.3.2 of Pessoa et al. 1998a).

The results of several studies suggest that the representation of certain amodally completed surfaces is "equivalent" to that of an associated "complete" version. In other words, some early stages of the visual system treat the two types of stimuli shown in Figure R1 as equivalent. This is evidenced by the slow search times when amodally-completed targets are searched in a field of distractors (Rensink & Enns 1998; see also Davis & Driver 1998; He & Nakayama 1994).

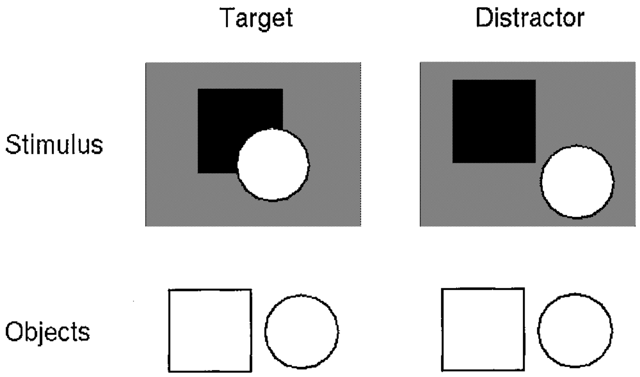

Figure R1 (Pessoa et al.). Amodal completion study by Rensink and Enns (1998). When subjects search for the target among a field of distractors, their search times are slow, indicating that they do not have access to the individual elements (notch plus circle), but instead to something that more closely resembles the distractors (hence the difficulty). Thus, the target and the distractor are treated similarly by the early visual system (indicated by the second row of "objects"). Note, however, that no evidence for the actual representation of a filled-in surface was obtained.

Other studies suggest that the visual system treats modal and amodal completion quite differently. Some recent studies of visual attention provide an example. It is well known that attention "spreads" within regular, non-illusory surfaces. For instance, subjects can identify two attributes of a single object more efficiently than two attributes of different objects (e.g., Baylis & Driver 1993). More generally, several well-known effects of attention have been found to apply to perceptual objects and not just spatial locations (for discussion and references, see Lavie & Driver 1996; Moore et al. 1998). Davis and Driver (1997; 1998) investigated whether attention would spread within modally and amodally completed illusory surfaces. (The two types of surfaces differed minimally in their physical attributes, but perceptually were quite different.) To their surprise, they found that the two types of surfaces behaved quite differently in relation to attention: modally completed surfaces exhibited the attention-spreading effect found in real, non-illusory surfaces, whereas amodally completed surfaces did not. As a result, Davis and Driver suggest that the perceptually salient, foreground properties of the modal surface might be related to processes of attention.

In the studies discussed in the previous two paragraphs, the central issue is not whether completion is perceptual or cognitive, but rather how the early visual system treats modally completed and amodally completed surfaces. The fact of poor search performance (when amodally completed targets are searched in a field of distractors) appears to indicate that normal surfaces and amodally completed surfaces are both "real" to vision. Nevertheless, modal and amodal completions have very different perceptual qualities: for example, the latter have a less "present" or "encountered" character than the former (Kanizsa & Gerbino 1982). If Davis and Driver (1997; 1998) are right, this difference may be associated with attention. We wish to stress that unless we dissect the functional effects of completion, we risk missing the subtleties of these complex phenomena.

We have argued that perceptual completion comprises a wide range of different phenomena, and therefore that it is important not to prejudge such issues as whether modal and amodal completion involve common neural mechanisms. Indeed, it seems likely that the completion of a cat behind a picket fence will depend on neural structures and circuits that are not implicated in the perception of Kanizsa figures. (There is evidence that the perceptual completion in Kanizsa figures depends on "lower-level" processes: see Ffytche & Zeki 1996; Hirsch et al. 1995). In any case, it seems clear that acknowledgment of the heterogeneity of different completion phenomena is necessary if we are to advance our understanding of their various neural substrates.

Vladusich accepts our position presented in the target article, in particular our criticisms of analytic isomorphism and our view that perceptual content needs to be understood at the level of the animal interacting with the world. He proposes that "resonant binding" among distributed brain regions may play a role in perceptual completion. Although this hypothesis is not unattractive, much more evidence is needed, especially evidence that addresses particular cases of perceptual completion. In general, however, we think that "dynamic brain mapping" of phase synchrony/desynchrony in neural assemblies is one of the most promising approaches to studying the neural basis of perception (see Rodriguez et al. 1999; Varela 1995), and we expect to see significant advances along this front in coming years.

ACKNOWLEDGMENTS
L. P. was supported in part by grants from CNPq, PRONEX, and FUJB. E. T. acknowledges the support of the Social Sciences and Humanities Research Foundation of Canada. A. N. gratefully acknowledges the support of faculty research funds granted by the University of California, Santa Cruz.

References

Letters "a" and "r" appearing before authors' initials refer to target article and response, respectively.

Ballard, D. H. (1991) Animate vision. *Artificial Intelligence* 48:57–86. [rLP]

Barnes, J., Howard, R. J., Senior, C., Brammer, M., Bullmore, E. T., Simmons, A. & David, A. S. (1999) The functional anatomy of the McCollough contingent colour after effect. *Neuroreport* 10:195–99. [VT]

Baylis, G. C. & Driver, J. (1993) Visual attention and objects: Evidence for hierarchical coding of location. *Journal of Experimental Psychology: Human Perception and Performance* 19:451–70. [rLP]

Berkeley, J., Debruyn, B. & Orban, G. (1994) Illusory, motion, and luminance-defined contours interact in the human visual system. *Vision Research* 34:209–16. [rLP]

Broerse, J., Vladusich, T. & O'Shea, R. P. (1999) Colour at edges and colour spreading in the McCollough effect. *Vision Research* 39:1305–20. [VT]

Buffart, H., Leeuwenberg, E. & Restle, F. (1981) Coding theory of visual pattern completion. *Journal of Experimental Psychology: Human Perception and Performance* 7:241–74. [RvL]

Clark, A. (1996) *Being there: Putting brain, body, and world together again.* MIT Press. [rLP]

Davis, G. & Driver (1997) Spreading of visual attention across modally versus amodally completed surfaces. *Psychological Science* 8:275–81. [rLP]

(1998) Kanizsa subjective figures can act as occluding surfaces in preattentive human vision. *Journal of Experimental Psychology: Human Perception and Performance* 8:275–81. [rLP]

Dennett, D. C. (1991) *Consciousness explained.* Little Brown [RvL, rLP]

Ffytche, D. H. & Zeki, S. (1996) Brain activity related to the perception illusory contours. *Neuroimage* 3:104–108. [rLP]

Gibson, J. J. (1979) *The ecological approach to visual perception.* Houghton Mifflin. [VT]

Grossberg, S. (1980) Direct perception or adaptive resonance? *Behavioral and Brain Sciences* 3:385–86. [VT]

(1995) The attentive brain. *American Scientist* 83:438–49. [VT]

He, Z. J. & Nakayama, K. (1994) Perceiving texture: Beyond filtering. *Vision Research* 34:151–62. [rLP]

Hirsch, J., DeLaPaz, R. L., Relkin, N. R., Victor, J., Kim, K., Li, T., Borden, P., Rubin, N. & Shapley, R. (1995) Illusory contours activate specific regions in human visual cortex: Evidence from functional magnetic imaging. *Proceedings of the National Academy of Sciences (USA)* 92:6469–73. [rLP]

Kanizsa, G. (1985) Seeing and thinking. *Acta Psychologica* 59:23–33. [RvL]

Kranda, K. (1998) Blindsight in the blind spot. *Behavioral and Brain Sciences* 21:762–63. [rLP]

Lavie, N. & Driver, J. (1996) On the spatial extent of attention in object-based visual selection. *Perception and Psychophysics* 58:1238–51. [rLP]

Marr, D. (1982) *Vision.* W. H. Freeman. [VT]

Moore, C. M., Yantsis, S. & Vaughan, B. (1998) Object-based visual selection: Evidence from perceptual completion. *Psychological Science* 9:104–10. [rLP]

Neumann, H. (1998) Representations, computation, and inverse ecological optics. *Behavioral and Brain Sciences* 21:766–67. [rLP]

Noë, A & Thompson (1999) Seeing beyond the modules to the subject of perception. *Behavioral and Brain Sciences* 22:386–387. [rLP]

Noë, A. Pessoa, L. & Thompson, E. (2000) Beyond the grand illusion: What change blindness really teaches us about vision. *Visual Cognition* 7(1–3): 93–106. [rLP]

McCollough, C. (1965) Color adaptation of edge detectors in the human visual system. *Science* 149:1115–16. [VT]

Paradiso, M. A., Shimojo, S. & Nakayama, K. (1989) Subjective contours, tilt-aftereffects, and visual cortical organization. *Vision Research* 29:1205–13. [rLP]

Pessoa, L. & Neumann, H. (1998) Why does the brain fill in? *Trends in Cognitive Science* 2:422–24. [rLP]

Pessoa, L., Thomson, E. & Noë, A. (1998a) Finding out about filling-in: A guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences* 21:723–802. [RvL, TV, rLP]

Peterhans, E. & von der Heydt, R. (1989) Mechanisms of contour perception in monkey visual cortex. II. Contours bridging gaps. *Journal of Neuroscience* 9:1749–63. [rLP]

Pylyshyn, Z. (1999) Is vision continuous with cognition? The case for the cognitive impenetrability of vision. *Behavioral and Brain Sciences* 5(11):499. [rLP]

Ramachandran, V. S. (1992) Blind spots. *Scientific American* 266:86–91. [VT]

Rensink, R. A. & Enns, J. T. (1998) Early completion of occluded objects. *Vision Research* 38:2489–505. [rLP]

Rodriguez, E., George, N., Lachaux, J-P., Martinerie, J., Renault, B. & Varela, F. J. (1999) Perception's shadow: Long-distance synchronization of human brain activity. *Nature* 397:430–33. [rLP]

Sekuler, A. (1994) Local and global minima in visual completion: Effects of symmetry and orientatiion. *Perception* 23:529–45. [RvL]

Sekuler, A., Palmer, S. & Flynn, C (1994) Local and global processes in visual completion. *Psychological Science* 5:260–67. [RvL]

Thompson, E., Palacios, A. & Varela, F. J. (1992) Ways of coloring: Comparative color vision as a case study for cognitive science. *Behavioral and Brain Sciences* 15:1–74. [rLP]

Ullman, S. (1980). Against direct perception. *Behavioral and Brain Sciences* 3:373–415. [VT]

van Lier, R. (1999) Investigating global effects in visual occlusion: From a partly occluded square to the back of a tree trunk. *Acta Psychologica* 102:203–20. [RvL]

van Lier, R., Leeuwenberg, E. & Van der Helm, P. (1995a) Multiple completions primed by occlusion patterns. *Perception* 24:727–40. [RvL]

van Lier, R., Van der Helm, P. & Leeuwenberg, E. (1994) Integrating global and local aspects of visual occlusion. *Perception* 23:883–903. [RvL]

(1995b) Competing global and local completions in visual occlusion. *Journal of Experimental Psychology: Human Perception and Performance* 21:571–83. [RvL]

van Lier, R. & Wageman, J. (1999) From images to objects: Global and local completions of self-occluded parts. *Journal of Experimental Psychology: Human Perception and Performance* 25:1721–41. [RvL]

Varela, F. J. (1995) Resonant cell assemblies. A new approach to cognitive functions and neuronal synchrony. *Biological Research* 28:81–95. [rLP]

Varela, F. J., Thompson, E. & Rosch, E. (1991) *The embodied mind: Cognitive science and human experience.* MIT Press. [rLP]

von der Heydt, R. & Peterhans, E. (1989) Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. *Journal of Neuroscience* 9:1731–48. [rLP]

Walker, R. & Mattingley, J. B. (1998) Pathological completion: The blind leading the mind? *Behavioral and Brain Sciences* 21:778:79. [VT]

Whittle, P. (1998) Filling-in does require a mechanism, and some persistent doubts. *Behavioral and Brain Sciences* 21:779–80. [VT]

Wouterlood, D. & Boselie, F. (1992) A good-continuatiion model of some occlusion phenomena. *Psychological Research* 54:267–77. [RvL]

Commentary on **C. M. Heyes (1998) Theory of mind in nonhuman primates. BBS 21:101–148.**

Abstract of the original article: Since the BBS article in which Premack and Woodruff (1978) asked "Does the chimpanzee have a theory of mind?," it has been repeatedly claimed that there is observational and experimental evidence that apes have mental state concepts, such as "want" and "know." Unlike research on the development of theory of mind in childhood, however, no substantial progress has been made through this work with nonhuman primates. A survey of empirical studies of imitation, self-recognition, social relationships, deception, role-taking, and perspective-taking suggests that in every case where nonhuman primate behavior has been interpreted as a sign of theory of mind, it could instead have occurred by chance or as a product of nonmentalistic processes such as associative learning or inferences based on nonmental categories. Arguments to the effect that, in spite of this, the theory of mind hypothesis should be accepted because it is more parsimonious than alternatives or because it is supported by convergent evidence are not compelling. Such arguments are based on unsupportable assumptions about the role of parsimony in science and either ignore the requirement that convergent evidence proceeds from independent assumptions, or fail to show that it supports the theory of mind hypothesis over nonmentalist alternatives. Progress in research on theory of mind requires experimental procedures that can distinguish the theory of mind hypothesis from nonmentalist alternatives. A procedure that may have this potential is proposed. It uses conditional discrimination training and transfer tests to determine whether chimpanzees have the concept "see." Commentators are invited to identify flaws in the procedure and to suggest alternatives.

More theory and evolution, please!

Radu J. Bogdan
Department of Philosophy, Tulane University, New Orleans, LA 70118.
bogdan@mailhost.tcs.tulane.edu

Abstract: Heyes's (1998) skepticism about theory of mind (ToM) in nonhuman primates exploits the idea of a strong and unified theory of mind in humans based on an unanalyzed category of mental state. It also exploits narrow debates about crucial observations and experiments while neglecting wider evolutionary trends. I argue against both exploitations.

Heyes's (1998) thought-provoking target article is a reminder that a convincing argument for a theory-of-mind competence in nonhuman primates remains hard to elaborate and defend. Lack of robust naturalistic data and the rarity of decisive experiments are not the only reasons, although they are reasons perhaps best exploited by Heyes's skepticism. Equally frustrating may be some lack of conceptual clarity and the hesitance to theorize beyond the narrow boundaries of data and experiments. It is the latter shortcomings, also exploited by Heyes, that I want to address briefly in what follows.

Why mental state? I am puzzled by the notion of mental state placed at the heart of most analyses of the ToM competence. What could it mean? And why treat it as a premise rather than an outcome of inquiry? Philosophers have long disagreed about what mental states are, cognitive scientists do not really care, ordinary language is rather opaque but resolutely flexible and pragmatic about it, and yet from early on many if not most ToM theorists have made the mastery of a naïve category of mental state a test of having the ToM competence. So does Heyes, using the test as a prop for her skepticism. I think this is misleading and will explain why in a moment. Right now I want to make a general methodological point.

It is a well-known metascientific axiom that theoretical concepts in science, such as gravitation or gene, are usually defined implicitly in a given theory in terms of basic laws, causal or functional implications, and various assumptions. It is a theory of the ToM competence which ought to determine the nature of the categories that structure and run the competence. It cannot be a prior and pretheoretical decision or expectation of what the competence should be. This applies also to the human ToM: we should not regard it as a genuine ToM because it fits pretheoretical expectations about mental state concepts but because its theoretical analysis suggests so. In short, the nature of the ToM categories and skills ought to follow from, not premise, the theoretical inquiry. I am afraid that a good part of the ToM field today suffers from a methodological inversion. This inversion is exploited by skeptics like Heyes. Two further and misleading assumptions also help such skepticism.

Unhelpful premises. Like many workers in this field, Heyes accepts two premises about ToM which are misleadingly modeled on the adult human competence. One premise is that of a *strong* ToM, based on mental state categories, such as belief, desire, knowledge, which pick up types of mental conditions with causal powers and often clear-cut verbal and behavioral manifestations. The other premise is that of a *unified* ToM that operates by smoothly integrating such concepts. I think these premises generate the wrong conditions for the argument for or against ToM in *nonhuman* primates. Not having the space to elaborate this diagnosis (see Bogdan 1997; 1999) I restrict myself to a few pointers.

The first premise first. A ToM competence need not be mentalist (i.e., based on mental state concepts that capture internal conditions with causal powers) in order to be made of specialized, functionally dedicated, and domain-specific skills. This is why an argument for (or against) ToM must first reveal (or rule out) such skills, *not* their mentalist character. If it turns out that nonmentalist skills are actually involved in the apes' recognition of emotion, information access (seeing), and gaze following, then there is a viable alternative to the exclusive (and probably wrong) disjunction between a mentalist ToM and associative learning (see also Gordon 1998).

The apes' inability to recognize knowledge or attention in gaze or seeing accordingly, need not invalidate the existence of domain-specific and specialized skills; rather it clarifies their functional profile and limitations. Which brings me to the second premise, about a unified ToM. There is no good reason to believe that ToM is one tightly knit and homogenous competence or that it emerged wholesale at some discrete point in primate phylogeny or human ontogeny. After 20 years of intense research, there are optimistic grounds for believing that the primate ToM is made of a rich battery of skills that had evolved in fits and starts and in a variety of social, interpersonal, and cultural ambiances, building upon or converging with or inserting themselves in still other skills, and that only later in human childhood did language-based and thoroughly enculturated upgrades of such skills acquire a mentalist profile and job. These later acquisitions provide no grounds to downgrade its precursors to mere learning or chancy guesses, however.

Transcending the proximate. Yet neither of these points against Heyes's premises have much bite if we fail to take a wider theoretical and evolutionary view of the matter, something that Heyes does not do. Let me begin with a theoretical point that addresses Heyes's skepticism about ToM in apes. Most organisms evolve specialized organs and skills *under assumptions* about the ecology in which they operate. (To cite major instances, Vogel 1988 analyzes this phenomenon in biology, in general, and Marr 1982 specifically in vision.) The body design of fish and birds evolved under assumptions about the properties of the environments they travel through (water, air); on the artificial side, engineers would design ships and planes under rather similar assumptions. Assumptions are not an explicit part of the design of an organ or skill or artifact (so they cannot be determined by just looking at design) but are part of a wider ecology-organ/skill/artifact package, which is naturally selected or deliberately envisaged *as a whole.* The same is true of *cognitive* organs and skills. The visual system works under assumptions about light bouncing off surfaces, boundaries revealing shapes and volumes, and the like.

As part of cognition, the ToM skills are no exception. Assumptions about ecologies, natural and social, are an indelible complement to their design and operation. Assumptions also lead to a sort of *division of labor* between a skill's reach and the ecology's completion of it: a skill usually exploits regularities or landmarks to do what it was designed to do (Bogdan 1994). Thus, it could be that the sociopolitical ecology of apes calls for specialized ToM skills to track gaze and its direction (given their variable manifestations and utilities and the urgency of their manipulation) but not its targets (given that what one gazes at is a basic goal shared by everybody or can be determined contextually from various clues). It could also be that the job of a ToM skill in apes or human children consists precisely in patterning the right ecological, bodily, and behavioral clues to guide a response behavior. That would still make the still domain-specific and functionally dedicated, but without any mentalist import. In general, if ecology-sensitive assumptions and divisions of labor are not factored into the analysis of a ToM skill, it may be hard to devise experiments that reveal its nature, particularly when assumptions or divisions of labor or both are violated, as they may well be in laboratory contexts.

An evolutionary perspective, absent from Heyes's target article, would do more than reveal assumptions and divisions of labor. It could also suggest, perhaps better than observation and experiment, whether there were selective pressures and opportunities for a specialized ToM competence. Such a suggestion would not settle the matter but would most likely constrain and inform the theoretical expectations. An analysis of the genetic proximity between apes and humans could further narrow the estimates. As important and complementary would be an inquiry into the maturational schedule of the presumed ToM skills. What convinced many researchers that humans develop a specialized ToM competence was the tight ontogenetic scheduling of some of its key skills. Such a schedule suggests genetic expression and the latter in turn suggests some evolutionary pedigree. I would accordingly be advisable to see whether there is such an ontogenetic scheduling in *ape* infancy and childhood as well, particularly for skills, such as gaze recognition and gaze following, which many researchers regard as domain-specialized and probably inherited by humans.

Finally, a look at the evolution of ToM skills in primates may also shed light on the interpersonal, cultural, and linguistic pressures and opportunities that eventually led to mentalist ToM skills in humans. Whatever the form of these skills (modular, naively theoretical, inferential), the fact that they had to factor in mental conditions with causal powers must be explained (instead of assumed) and so must the very possibility of this factoring. It is rather unlikely that such a possibility sprang into existence out of nowhere, without domain-specific and dedicated precursors that reach back into primate phylogeny. Nor should it be any surprise (should it?) that these precursor skills do not look human and do not operate as they do in humans.

Theory of mind and the "somatic marker mechanism" (SMM)

Bruce G. Charlton

Department of Psychology, University of Newcastle upon Tyne, NE1 7RU, United Kingdom, **bruce.g.charlton@ncl.ac.uk**
www.hedweb.com/bgcharlton/

Abstract: The "somatic marker mechanism" (SMM; Damasio 1994) is proposed as the cognitive and neural basis of the theory of mind mechanism. The SMM evolved for evaluating the intentions, dispositions, and relationships of conspecifics; hence, it is adaptive in the social domain. It is predicted that chimpanzees will indeed have theory of mind (ToM) ability, but that this will be socially domain-specific. Domain-*general* ToM will be found only in primates with abstract, symbolic language (adult humans). Putative ToM tests require revision in the light of these distinctions.

The nature of the putative "theory of mind mechanism" (ToMM), and whether or not it extends beyond the human species, has created great controversy among evolutionary biologists and animal behaviorists – most recently in the form of Heyes's 1998 *BBS* target article and the accompanying comments. Unfortunately, the discussion of this topic omits what should be regarded as the key neuroscientific references relevant to this topic: the work by Damasio and colleagues on the "somatic marker mechanism (SMM) (Damasio 1994; 1995; 1996). I suggest that the SMM forms the cognitive and neural basis of the ToMM (see Premack & Woodruff 1978).

The most commonly used "pure cognitive" conceptualization of ToM defines the mechanism in terms of a capacity to represent the *contents* of other minds – that is, to represent their distinctive mental states. This implies that ToM uses a two-fold cognitive representation: a representation of the mind of a conspecific, and the contents of that mind (leading to those witty cartoons of one "thought balloon" inside another). The pure cognitive conceptualization of ToM sees the representational mechanism as abstract, symbolic, and domain-general.

The special qualities of "intelligence" seen in primates, however, are more plausibly seen as domain-specific and driven principally by the demands of social living (Byrne & Whiten 1988; Dunbar 1996). Hence, ToM should be conceptualized as a mechanism which evolved for, and is adaptively concerned with, understanding, predicting, and manipulating the behaviour of conspecifics. The theory of mind mechanism is specialized for representing social variables such as dispositions, intentions, and relationships.

Theory of mind can be seen as the means by which *overt* behaviour is interpreted in the light of *inferred mental attributes.* Most animals, lacking a ToMM, infer the meaning of behaviour directly from overt behavioural cues. But in an animal using ToM cognition, the primary interpretative inference is "mentalistic," and overt behaviour is understood in the context of an ascribed state of mind (e.g., Charlton & Walston 1998).

The selection pressure which led to the evolution of ToM was probably the potential ambiguity of social behaviour when overt behaviour is ambiguous (e.g., when behaviour is complex, multivalent, rapidly changing, or deceptive; Byrne & Whiten 1988; Dunbar 1996). When cues are ambiguous, interpretation of a given cue becomes dependent upon inferences concerning intentions, dispositions, and relationships. For example, the approach of another human may have several meanings. In the interpretative sequence "he is angry, and approaching me – therefore I must get ready to fight"; the mentalistic ascription of anger is logically prior to the interpretation of overt behavioral cues. If the ascription of disposition were to be changed from "angry" to "happy," then – even when the immediately perceived cues are identical – the inferred meaning of the overt behaviour "approaching me" (and the implications for an adaptive response) would also change.

The work of Damasio (1994; 1995; 1996) and colleagues is crucial to understanding ToM because it provides an integrated functional and neural explanatory model of this behavioral sequence and it defines the adaptive scope of the mechanism. Emotions and feelings are brain representations of physiological states; these reach awareness when projected to working memory as topographically organized patterns of neural activity. Fluctuations of the inner bodyscape (i.e., emotional responses) form the means by which social sensory inputs are evaluated and strategic social modeling is performed. The somatic marker mechanism (SMM) therefore refers to the process by which changes in the *soma* (body) are used to *mark* perceptual inputs when somatic and perceptual representations are temporally juxtaposed in working memory. And the class of perceptions in question relate specifically to *social* situations since ToM evolved as an aspect of social intelligence and to solve social problems.

As a plausible example in chimpanzees, a male stranger might evoke the "fight or flight" response. This response comprises a characteristic physiological state driven by the sympathetic nervous system (vasoconstriction in skin, vasodilatation in muscles, hairs standing on end, sweating, faster heart rate, etc.). Cognitive representations of changing body state are continually constructed in the brain from feedback from the afferent nerves, chemo-receptors, and other inputs. In an animal lacking a ToMM, such cognitive representations of the changing body state may affect behaviour – perhaps by provoking involuntary flight. But in an animal with an SSM, a cognitive representation of this changing body state may be projected to working memory (WM) where it becomes accessible to awareness as conscious fear. The cognitive representation of "fear" may then be used as a somatic marker when sustained in WM in temporal juxtaposition to the perceptual representation of the male stranger's identity.

The juxtaposition of the somatic marker for fear with the stranger's identity that evoked it is assumed to create a novel cognitive representation incorporating what is, in effect, the *disposition* of that individual. The combined representation is implicitly one of "that fear-evoking stranger": that is, aggression and hostility are attributed as a "theory" of the stranger's mental contents. The combined social identity/body-state representation can be stored in long term memory, and when recalled to WM it will be capable both of recollecting individual identity and re-enacting the linked emotion of fear as a change in body state. Hence, social identities and their relationships can be modeled (variously combined and sequenced), and the consequences of this modeling evaluated (as gratifying or aversive) by re-experiencing the enacted emotional body state. Somatic marking is therefore proposed as the actual mechanism of ToM, and in this sense the SMM is the basis of "mindreading" (to use Baron-Cohen's 1995 term): the SMM is a mechanism for inferring what are *de facto* intentions and dispositions. Nonetheless, the SMM could be considered almost the *reciprocal* of the common cognitive conceptualization of the ToMM. For example, hostility would not be represented directly as the hostile contents of another's mind, but instead as the reciprocal attribution of the feeling of "fear." A "hostile" animal would actually be represented by the SMM as a "fear-evoking" animal – an identity "marked" by emotion.

I am arguing that the SMM will be found to comprise the basis of the theory of mind mechanism in young children, chimpanzees, and perhaps other nonhuman primates. However, in the existing ToM literature, the essential features of the ToM mechanism have been conflated with features that are actually attributes of abstract symbolic language, and therefore unique to language-using adult humans. Language is necessary for the two-fold representation (cartoon double "thought balloon") of other minds, and of their contents. Hence, language provides extra capabilities for adult human ToM by allowing the representation of domain-general, nonsocial knowledge.

The SMM theory predicts that chimpanzees will indeed have "theory of mind"; but only in the social domain. This ToMM comprises an ability to represent and model the dispositions, intentions, and social interactions of conspecifics. Because they lack an abstract, symbolic language, chimpanzees would *not* be expected to display a domain-general capacity to represent the nonsocial "contents" of others' minds when these extend beyond the social domain.

Many of the putative "ToM tests" which have been used in children and primates do not distinguish social and nonsocial domains. Instead, they demand, for instance, inferences about knowledge of spatial location, which lies beyond the scope of the SMM. This failure to consider ToM separately in non-language users, has probably served to perpetuate controversy over the species-distribution of ToM. "ToM tasks" need to be re-designed to measure SMM-specific reasoning processes. Only such socially specific ToM tests can answer the question of how far ToM extends beyond humans throughout primate species, and whether it is found in other social mammals such as elephants and dolphins.

NOTE ADDED IN PROOFS

The above ideas have since been expanded and refined in a book, *Psychiatry and the human condition* (2000).

How to solve the distinguishability problem: Triangulation without explicit training

Robert W. Lurz

Department of Social Sciences, Indian River Community College, Ft. Pierce, FL. **rlurz@ircc.cc.fl.us**

Abstract: Heyes's (1998) triangulation approach to distinguishing a "theory" of mind (ToM) from a "theory" of behavior (ToB) in chimpanzees fails. The ToB theorist can appeal to the explicit training sessions and analogical reasoning to explain/predict the chimpanzees' behaviors. An alternative triangulation experiment is sketched, demonstrating how the removal of such training sessions paves the way toward solving the distinguishability problem.

There are two rival hypotheses about how nonhuman primates predict/anticipate the behaviors of other animals. The "theory" of mind hypothesis (ToM) maintains that such predictions/anticipations are mediated by mental-state attributions (e.g., Premack & Woodruff 1978). The "theory" of behavior hypothesis (ToB) maintains that such predictions/anticipations are mediated by associative or inferential learning about stimulus-response contingencies (e.g., Heyes 1993). A number of theorists have charged that continued research into ToM in animals faces an intractable methodological problem: It appears that any experimental datum that can be plausibly explained/predicted by a ToM hypothesis – by assuming the animal in question takes stimulus S to be correlated with mental state M and M with behavior R in another animal – can be plausibly explained/predicted by a ToB hypothesis – by assuming that the animal in question simply learns that S is correlated with R in another animal.

In the target article, Heyes (1998) attempts to solve this *distinguishability problem* by designing an experiment whose positive results cannot plausibly be explained/predicted by assuming that the test animals (chimpanzees) learn that a stimulus (eye-object line) is correlated with a particular behavior (pointed to the baited containers) in the trainers, but can only be explained/predicted by a ToM hypothesis. Unfortunately, the positive results of Heyes's experiment can be explained/predicted by a ToB hypothesis that allows analogical reasoning in chimpanzees. According to this hypothesis, the chimpanzees come to learn that there is a correlation between (S*) wearing the red-trimmed goggles with transparent lenses in front of an object and (S) the object being directly before their eyes. After all, it's reasonable to suppose that while wearing the red-trimmed goggles before some object *x*, the chimpanzees saw that *x* was directly before their eyes. Armed with this information, the chimpanzees could very well choose the Knower on the probe trials by choosing the trainer who had direct eye contact with the baiting process. For the chimpanzees could reason that since there was a correlation in their own cases between (S*) wearing the red-trimmed goggles in front of an object and (S) the object being directly before their eyes, the same S*–S correlation holds between the Knower's eyes and the baiting process. To head off a possible objection here, it should be noted that (1) there is suggestive evidence that chimpanzees are capable of analogical reasoning (Gillian et al. 1981), and (2) positing such a capacity in chimpanzees is quite consistent with the ToB hypothesis.

The mistake in Heyes's triangulation experiment is the use of explicit training prior to the probe trials that allows the chimpanzees to associate/infer that observable cue S (eye-object line) is correlated with behavior R (pointing to the baited container) in the trainers. I believe that once this type of training is removed, a path is cleared toward solving the distinguishability problem. To illustrate, consider the following triangulation experiment *sans* explicit training:

(1) Pretraining stage: the chimpanzees are allowed to examine two pairs of goggles – red-trimmed goggles with translucent lenses and blue-trimmed goggles with opaque lenses.

(2) Training stage: the chimpanzees are presented with four containers, one of which is baited by a third trainer while the Knower and the Guesser remain in the room. The baiting process is done behind a screen so that neither the chimpanzee *nor* the Knower or the Guesser sees which container is baited. After the screen is removed, the Knower and the Guesser point to a different container. The chimpanzees are trained to choose one of these containers. If the container is baited, then the trainer and the chimpanzee receive a reward while the other trainer does not; if the container is not baited, then neither the trainer nor the chimpanzee receives a reward while the other trainer does, provided he chooses the baited container.

(3) Probe trial: During the probe trial, the chimpanzees observe that both the Knower (wearing the red-trimmed goggles) and the Guesser (wearing the blue-trimmed goggles) are directly facing the baiting process while standing behind the screen with the third trainer. The chimpanzees cannot see which container is being baited as a result of the placement of the screen.

In this experiment, ToM and ToB have different predicted outcomes. Since the chimpanzees are not taught in the training trial that an observable cue, such as eye-object line or red-trimmed goggles, is correlated with a particular behavior in the trainers, such as pointing to baited containers, ToB predicts that the chimpanzees will choose randomly. However, ToM predicts that the chimpanzees will favor the Knower over the Guesser. For, according to the ToM hypothesis, the chimpanzees learned that wearing red-trimmed goggles in front of an object is correlated with *seeing* the object and inferred that the same correlation holds in the Knower's case. Furthermore, ToM can suppose that during the training trials, the chimpanzees learned that each trainer *wanted* to be rewarded and *believed* that he would be if he chose the baited container. A ToM theorist can postulate, as some have (e.g., Segal 1996), that these mental-state attributions are plugged into a ToM module that is partly defined by intentional laws, such as: if one sees *p*, the (*ceteris paribus*) one knows *p;* and if one knows *p*, wants *r*, and believes that one will get *r* if one does *q* when *p* is the case, then (*ceteris paribus*) one will do *q*. From this, the chimpanzees will predict that the Knower will point to the baited container. Although this solution to the distinguishability problem assumes a modularity thesis about animal ToM, I do not see that this should prevent it from being a successful solution.

Author's Response

Theory of mind and other domain-specific hypotheses

C. M. Heyes

Department of Psychology, University College London, London WC1E 6BT, United Kingdom. **c.heyes@ucl.ac.uk**
www.psychol.ucl.ac.uk/celia.heyes/netintro.html/

Abstract: The commentators do not contest the target article's claim that there is no compelling evidence of theory of mind in primates, and recent empirical studies further support this view. If primates lack theory of mind, they may still have other behavior control mechanisms that are adaptive in complex social environments. The Somatic Marker Mechanism (SMM) is a candidate, but the SMM hypothesis postulates a much weaker effect of natural selection on social cognition than the theory of mind hypothesis (on inputs to cognitive mechanisms, not on the mechanisms themselves), and there is currently no evidence that it is specific to social stimuli or to primates. "Two Guesser" training would make the goggles test too chauvinistic, and in its current form the goggles problem could not be solved by physical matching because, while wearing goggles, an individual cannot see itself seeing.

Bogdan and **Charlton** apparently agree with the main thrust of my target article, that there is no compelling evidence of theory of mind in primates. However, they each emphasize that if primates lack theory of mind as it is conceptualized by those conducting empirical research with primates (and therefore also in the target article), it does not necessarily mean that primates lack domain-specific social cognitive abilities. Although this is obviously true, Bodgan's and Charlton's commentaries are valuable because in making the point they raise important issues relating to the formulation and evaluation of hypotheses about the evolution of social cognition. I will turn to these issues after responding to **Lurz** who, not yet ready to give up on theory of mind in primates, suggests a modification to my goggles experiment.

R1. Two Guessers

Lurz suggests that during the training phase of the goggles experiment the chimpanzees should be required to choose between two containers, one indicated by each of two trainers who are both guessing the location of the food because their view of the baiting process was blocked by a screen. This "Two Guessers" proposal contrasts with my "Knower-Guesser" proposal (Heyes 1998, Response) that, during training, chimpanzees choose between two containers, one (which contains the food) indicated by a trainer who saw the baiting, and another chosen at random by a trainer who did not have visual access to baiting. In the Knower-Guesser scenario, the chimpanzee is rewarded at the end of every trial in which it selects the container indicated by the Knower, and, on average, at the end of one in four trials in which it selects the container indicated by the Guesser. In the Two Guesser scenario, presumably, the chimpanzee would receive food at the end of one in four trials on average, regardless of the trainer he selects to use as a cue.

For three reasons, I suspect that the Two Guesser scenario would make the goggles experiment too chauvinistic, that is, at high risk of yielding a false negative result. First, the subjects may learn in the training phase that it does not matter which trainer they choose, and this could transfer to the probe trials, making them inattentive to which trainer is wearing the red and which the blue goggles. Second, in the Two Guesser arrangement subjects would be given no hint during training that the problems set in this experiment can be solved by discriminating between trainers according to whether they could see a critical event. Thus, chimpanzees that have the capacity to attribute sight may fail to use it on probe trials because they do not realize it is relevant to the test. Finally, if, as **Lurz** suggests, the subjects have an opportunity during Two Guesser training to learn "that each trainer wanted to be rewarded and believed that he would be if he chose the baited container" *and* if this information were necessary for successful probe trial performance, it would make the goggles experiment inappropriately complex, converting it from a test of attribution of seeing, to a test of attribution of belief, desire, and sight.

R2. Physical matching versus analogical reasoning

These considerations would not provide sufficient reason to use Knower-Guesser training if **Lurz** were right in suggesting that, as the goggles experiment stands, Knower preference on probe trials could be due to analogical reasoning based on the chimpanzee seeing itself seeing while it is wearing the red goggles at pretest. But when I, or a chimpanzee, put on the red, translucent goggles, I see what is before me; I do not, as if from a combination of first and third person perspectives, see myself wearing the red goggles, the object before my eyes, and myself seeing that object. This distinction is critical because, if the third person perspective were afforded by my wearing the goggles, Knower preference on the probe trials could be due to physical matching of stimuli. In other words, the chimpanzee may prefer the trainer wearing the red goggles over the trainer wearing the blue because the former looks more like the chimpanzee itself looked when it was wearing the red goggles and its reaching movements were effective, than like the chimpanzee itself looked when it was wearing blue and its reaching movements were ineffective. It was to prevent this kind of physical matching that I stressed in the target article that subjects should not see others wearing the red and blue goggles before the probe trials.

I assume that the goggles task requires analogical reasoning for its solution, but the critical analogy is between the chimpanzee's visual experience of the objects before it while wearing the red goggles during pretraining (e.g., cage walls, toys, its own limbs), and the chimpanzee's view of the trainer wearing red goggles on probe trials. The first of these would be no more physically similar to the sight of the trainer in red goggles than to the sight of the trainer in blue goggles. Therefore, Knower preference on probe trials would seem to require, not physical matching, but analogical reasoning along the lines: When I'm wearing the red goggles, I see what is before me. That trainer is wearing the red goggles, therefore he sees what is before him.

R3. Recent evidence

In agreement with **Lurz** and many others, I do not think that the evidence that primates lack a theory of mind has yet accumulated to the point where it is a waste of time to conduct well-designed experiments in this area. However, it is notable that the major experimental studies published since the target article have either reported negative findings (e.g., Call & Tomasello 1999; Reaux et al. 1999) or acknowledged that the demonstrated social competence is equally explicable in mentalistic and nonmentalistic terms (e.g., Hare et al. 2000). Call and Tomasello (1999) found no evidence of false belief attribution in a mixed group of orangutans and chimpanzees tested using a non-verbal version of the Sally-Anne task that had been validated with 4- and 5-year old children. It is unfortunate that the apes in this study had received extensive prior training in which a marker stimulus, that had to be avoided in the false belief choice tests (S-), functioned as a positive cue (S+). It would be advisable for any future studies using the same design to avoid this potential source of bias, and to ensure that the subjects cannot learn over test trials that the marker is an S-. However, as Call and Tomasello said, their false belief study provides absolutely no encouragement for the view that nonhuman apes can mentalize.

Hare et al. (1999) seem to present a more positive message, entitling their article "Chimpanzees know what conspecifics do and do not see," but in the discussion they acknowledge that their findings could be explained by a variety of behavioral and cognitive nonmentalistic hypotheses, as well by the hypothesis that chimpanzees attribute sight. In the terms used in the target article, Hare et al. show, at most, that chimpanzees can use "eye-object" line as a cue in competitive feeding situations, that is, that their choice of food targets can be influenced by whether or not there is or has been an unobstructed line between a competitors' eyes and the food object. That's smart, but, as Hare et al. firmly emphasize, it's not theory of mind – at least not as "theory of mind" has been conceptualized by developmental psychologists and primate researchers.

R4. Theory of mind and other domain-specific hypotheses

Bogdan characterizes the current conception of theory of mind as "strong," "unified," and "naïve," and argues that it was a mistake to drag it, largely unchanged, from the realm of day-to-day life into science. Instead of asking whether primates have, or when children begin to use, psychological resources defined by folk psychology or common sense, we should allow the products of research (theoretical and empirical) to inform a theory of the psychological resources involved in the prediction and explanation of behavior. I agree almost entirely with this argument, and if the target article provides Bogdan with ammunition in this battle – with examples of the vulnerability of research built on folk theory rather than scientific theory – I will be delighted. However, I say that I agree *almost* entirely because I would take issue with three of Bogdan's points.

First, **Bogdan** implies that I cast the theory of mind debate as one in which primates will either be found to have a strong, unified theory of mind, or shown to base all of their social behaviour on associative learning. In fact, I have explicitly rejected such a dichotomy, in the target article (sect. 2) and elsewhere (e.g., Heyes 1993).

Second, like **Charlton**, **Bogdan** seems to be ready to label as "theory of mind" any psychological resources that are found to underlie social competence in primates, or at least any such resources that are domain-specific – involved primarily or exclusively in social interaction. This is likely to breed confusion, with some people assuming that "theory of mind" refers to the "strong, unified" (Bogdan) or "abstract, symbolic" (Charlton) psychological resources to which it has referred over the last 20 years, and others using the term to indicate any psychological processes mediating social interaction.

My final disagreement with **Bogdan** is more substantial than either of the others, and it emphasizes some virtues of the "strong, unified" theory of mind hypothesis. Over the last 25 years, at least since the "social function of intellect hypothesis" (Humphrey 1976) was published, there has been a great deal of broad brush theorizing about the evolution of social cognition. What we need to do now is to convert plausible stories about the adaptive advantages of domain-specific social cognitive mechanisms into clearly specified, testable hypotheses and to subject them to empirical evaluation.

Like **Bogdan**, I think it is unfortunate that the theory of mind hypothesis came directly from folk psychology, but it has two significant virtues: it is a strong hypothesis in that it claims that social cognition is based on a distinctive psychological mechanism, not merely that is has distinctive input, and it is testable. If, as Bogdan and **Charlton** suggest, the focus of empirical research on social cognition in primates should now shift to other potentially domain-specific phenomena, the new hypothesis or hypotheses should have the same virtues, and if they postulate a weaker kind of domain-specificity (e.g., Suddendorf & Whiten 2001), it would be as well to recognize that retreat.

The somatic marker mechanism (SMM) hypothesis, favoured by **Charlton**, is clearly specified, interesting, testable, and it may well be true, but it makes much weaker claims than the theory of mind hypothesis about the effects of natural selection on (social) cognition. As it has been formulated by Damasio and his colleagues (e.g., Adolphs et al. 2000; Damasio 1996), the SMM hypothesis proposes that humans learn associations between exteroceptive input and somatic events characteristic of emotional response, and that these associations, which are stored in somatosensory cortices, modulate reasoning, and decision making. Thus, the hypothesis does not postulate any domain-specific psychological *mechanisms* (only associative learning, reasoning, and decision making), and while SMM associative learning has domain-specific *input,* its domain consists, not of social interaction, but of all environmental objects and events that provoke emotional responses. Furthermore, because a broad range of species show emotional responses, encounter "ambiguous" cues (Charlton), and are capable of associative learning, there is little reason (and currently no evidence) to suppose that the SMM is specific to primates.

It is beginning to look as if nonhuman primates do not have a theory of mind, but of course this does not mean that they lack behavioural control processes that are adaptive in complex social environments. Indeed, the existence of such processes is almost inevitable. The interesting questions are (Heyes 2000): Have these processes acquired their adaptive qualities phylogenetically or ontogenetically, and do they include domain-specific cognitive mechanisms, in addition to general-purpose cognitive mechanisms with domain-specific inputs?

References

Letters "a" and "r" appearing before authors' initials refer to target article and response respectively.

Adolphs, R., Damasio, H., Tranel, D., Cooper, G. & Damasio, A. R. (2000) A role for the somatosensory cortices in the visual recognition of emotion as revealed by three-dimensional lesion mapping. *Journal of Neuroscience* 20:2683–90. [rCMH]

Baron-Cohen, S. (1995) *Mindblindness: An essay on autism and theory of mind.* MIT Press. [BGC]

Bogdan, R. J. (1994) *Grounds for cognition.* Erlbaum. [RJB]

(1997) *Interpreting minds.* MIT Press. [RJB]

(1999). *Minding minds.* MIT Press. [RJB]

Byrne, R. W. & Whiten, A. (1988) *Machiavellian intelligence social expertise and the evolution of intellect in monkeys, apes and humans.* Clarendon Press. [BGC]

Call, J. & Tomasello, M. (1999) A nonverbal false belief task: The performance of children and great apes. *Child Development* 70:381–95. [rCMH]

Charlton, B. G. (2000) *Psychiatry and the human condition.* Radcliffe Medical Press. [BGC]

Charlton, B. G. & Walston, F. (1998) Individual case studies in clinical research. *Journal of Evaluation in Clinical Practice* 4:147–55. [BGC]

Damasio, A. R. (1994) *Descartes' error: Emotion, reason and the human brain.* Macmillan. [BGC]

(1995) Towards a neurobiology of emotion and feeling: Operational concepts and hypotheses. *Neuroscientist* 1:19–25. [BGC]

(1996) The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London B.* 351:1413–30. [BGC, rCMH]

Gillan, D. J., Premack, D. & Woodruff, G. (1981) Reasoning in the chimpanzee. I. Analogical reasoning. *Journal of Experimental Psychology* 7:1–17. [RWL]

Gordon, R. M. (1998) The prior question. *Behavioral and Brain Sciences* 21:120–21. [RJB]

Hare, B., Call, J., Agnetta, B. & Tomasello, M. (2000) Chimpanzees know what conspecifics do and do not see. *Animal Behaviour* 59:771–85. [rCMH]

Heyes, C. M. (1993) Anecdotes, training, trapping and triangulating: Do animals attribute mental states? *Animal Behavior* 46:177–88. [RWL, rCMH]

(2000) Evolutionary psychology in the round. In: *Evolution of cognition,* eds., C. M. Heyes & L. Huber. MIT Press. [rCMH]

Humphrey, N. K. (1976) The social function of intellect. In: *Growing points in ethology,* eds., P. P. G. Bateson & R. A. Hinde. Cambridge University Press. [rCMH]

Marr, D. (1982) *Vision.* W. H. Freeman. [RJB]

Premack, D. & Woodruff, G. (1978) Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1:515–26. [RWL]

Reaux, J. E., Theall, L. A. & Povinelli, D. J. (1999) A longitudinal investigation of chimpanzees' understanding of visual perception. *Child Development* 70:275–90. [rCMH]

Suddendorf, T. & Whiten, A. (2001) Mental evolution and development: Evidence for secondary representation in children, great apes, and other animals. *Psychological Bulletin* 127: 629–50. [rCMH]

Segal, G. (1996) The modularity of the theory of mind. In: *Theories of theories of mind.* P. Carruthers & P. K. Smith, eds. Cambridge University Press. [RWL]

Vogel, S. (1988) *Life's devices.* Oxford University Press. [RJB]

Commentary on **Phillipe G. Schyns, Robert L. Goldstone, and Jean-Pierre Thibaut (1998) The development of features in object concepts. BBS 21:1–54.**

Abstract of the original article: According to one productive and influential approach to cognition, categorization, object recognition, and higher lever cognitive processes operate on a set of fixed features, which are the output of lower level perceptual processes. In many situations, however, it is the higher level cognitive process being executed that influences the lower level features that are created. Rather than viewing the repertoire of features as being fixed by low-level processes, we present a theory in which people *create* features to subserve the representation and categorization of objects. Two types of category learning should be distinguished. Fixed space category learning occurs when new categorizations are representable with the available feature set. Flexible space category learning occurs when new categorizations cannot be represented with the features available. Whether fixed or flexible, learning depends on the featural contrasts and similarities between the new category to be represented and the individual's existing concepts. Fixed feature approaches face one of two problems with tasks that call for new features: If the fixed features are fairly high level and directly useful for categorization, then they will not be flexible enough to represent all objects that might be relevant for a new task. If the fixed features are small, sub-symbolic fragments (such as pixels), then regularities at the level of the functional features required to accomplish categorizations will not be captured by these primitives. We present evidence of flexible perceptual changes arising from category learning and theoretical arguments for the importance of this flexibility. We describe conditions that promote feature creation and argue against interpreting them in terms of fixed features. Finally, we discuss the implications of functional features for object categorization, conceptual development, chunking, constructive induction, and formal models of dimensionality reduction.

Feature development, object concepts, and the scope slip

Michael R. W. Dawson

Biological Computation Project, Department of Psychology, University of Alberta, Edmonton, Alberta, Canada T6E 2P9.

mdawson@psych.ualberta.ca www.bcp.psych.ualberta.ca/~mike

Abstract: Schyns et al.'s (1998) target article raises a conflict between the need for a fixed functional architecture in an explanatory cognitive science and the need for a system to learn to detect new features. This conflict can be resolved by avoiding the scope slip in which properties of objects are erroneously viewed as being properties of their representations.

Cognitive theories are usually "software" accounts: They describe some "mental program" that runs on the mind. Such theories are attractive because they are functional – they avoid making any detailed claims about the hardward (e.g., the brain) that is actually executing the "mental program."

Nevertheless, the explanatory power of a functionalist theory is rooted in the properties of physical mechanisms. Cognitive theories must ultimately be based upon a set of primitives whose performance is subsumed by natural, physical laws (e.g., Cummins 1983). In cognitive science, the proposed set of primitives for "mental programs" is called the *functional architecture* (e.g., Pylyshyn 1984). If a functional architecture is not proposed (and validated) for a cognitive theory, then this theory is merely a description, not an explanation. This is because in the absence of a set of primitives, a functional theory falls victim to the well-known homunculus problem (e.g., Edelman 1992, pp. 211–52).

One of the more intriguing themes running through the target article is the notion that adequate theories of object recognition or classification cannot be based upon a fixed functional architecture. Instead, Schyns et al. (1988) argue that the architecture must evolve according to the ongoing needs of a organism embedded in a particular environment. This approach is interesting because it has some alarming implications concerning the explanatory power of cognitive theories.

Specifically, can one claim that a functionalist theory counts as an explanation if its primitives are not fixed? For example, consider a situation in which a particular theory, based upon a fixed set of primitives, is criticized because it is not capable of recognizing an object that is characterized by a novel feature. One response to this criticism is to add to the existing model the capability of inventing a new primitive to represent the novel feature at issue. It is not clear that his revised theory is explanatory. For example, it is not at all obvious that such a theory would be falsifiable, particularly if new critiques could also be dealt with by inventing new primitives.

The conflict between fixed and varying functional architectures may emerge in the target article because Schyns et al. are the victims of what Pylyshyn (1981, p. 18) has called the *scope slip.* Pylyshyn introduced the scope slip as part of the imagery debate. He argued that the phrase *image of object X with property P* should be correctly interpreted as meaning *image of (object S with property P).* The scope slip involves a changed parsing of the phrase which results in a markedly different and incorrect interpretation: *(image of object X) with property P.* In other words, when the scope slip is committed, the properties of objects are erroneously viewed as being properties of the underlying representation (see also Pylyshyn).

With respect to the target article, the scope slip can be characterized as follows: Many cognitive theories of object recognition are concerned with providing accounts of the *representation of object X with feature F.* In some cases, *feature F* is a novel property – perhaps the system has to learn to use this new feature, as the evidence in the target article would suggest. Schyns et al. would have us believe that this learning results in a new architectural component, the (*representation of object X*) with *feature F.* However, it is much more likely that what is really at issue is the representation of (*object X with feature F*).

For example, in most parallel distributed processing (PDP) networks, the initial set of connection weights is randomly selected. As a result, prior to training, the network is unable to carry out a classification task of interest. As training proceeds, the performance of the network on the classification task improves. One reason for this improvement is that the network is actually learning about features of the stimuli. Indeed, and analysis of the internal structure of a trained network can reveal that it has discovered interesting features about the training set, and in some cases these features are both novel and psychologically interesting (Berkeley et al. 1995; Dawson et al. 1997).

Now the question is this: when PDP networks have learned to detect new features, are these features new components of the functional architecture? The answer to this question is that they are not. The functional architecture of a PDP network is the set of primitive network properties, including modifiable connections and the kinds of computations carried out by processing units (e.g., Dawson & Schopflocher 1992). When a PDP network learns a new feature, it is *not* building a new primitive. Instead, it is taking an existing set of primitives (weighted connections, processing units with a specific activation function) and organizing them into a feature detecting circuit. When the scope slip is not made, it is clear that the feature is a property of the stimuli that have been

0140-525X/01 $12.50

presented to the network, and the representation of that property is built from network components – not from features of the world.

By avoiding the scope slip, and by correctly treating features as properties of objects instead of properties of the architecture, it is possible to create an explanatory account of a system that evolves over time (e.g., by learning about new features). This is because the architectural account of this system would describe how new features emerge as representational properties. For example, an architectural account of a PDP network would explain how a learning rule, modifiable connections, and processing units would al interact to create circuits that could be interpreted as being feature detectors. Note that such accounts are falsifiable, because architectural accounts of PDP networks place strong constraints on what can and what cannot be learned (e.g., Minsky & Papert 1969/1988). In other words, by correctly treating features as being separate from the functional architecture, one should be able to make some claims about what new features could be learned, as well as about what new features could never be learned because they could never be represented by the functional architecture.

ACKNOWLEDGMENT
The writing of this paper was supported by NSERC Research Grant A2063.

Authors' Response

Functional identification of constraints on feature creation

Phillipe G. Schyns[a], Robert L. Goldstone[b], and Jean-Pierre Thibaut[c]
[a]*Department of Psychology, University of Glasgow, Glasgow, G12 8QB, United Kingdom;*[b]*Department of Psychology, Indiana University, Bloomington, IN 47405;* [c]*Department of Psychology, Université de Liège, Batiment B32, Sart-Tilman, 4000 Liège, Belgium.*
philippe@psy.gla.ac.uk www.psy.gla.ac.uk
rgoldsto@ucs.indiana.edu cognitrn.psych.indiana.edu/
jthibaut@ulg.ac.be

Abstract: Dawson's provocative comment makes three connected points: (1) to be falsifiable, theories that assume flexible features must constrain their feature creation and mechanisms, (2) the explanatory power of such functional theories is rooted in the properties of their underlying physical mechanisms, and (3) to derive the relevant constraints of feature creation from these mechanisms, it is critical to avoid the scope slip. We will argue here that even though we agree with (1) and (2), (3) confuses two different levels of analysis of computational systems: the functional *identification* and the physical *implementation* of relevant constraints.

R1. Constraints do matter

Dawson rightly points out that fixed feature theories can always be criticized, and falsified, whenever their postulated feature repertoire is not capable of recognizing an object characterized by a novel feature. In contrast, a flexible repertoire could, in principle, represent this novel feature with the creation of a new feature. A flexible, improperly constrained system might therefore be forever immune to falsification because it could deal with new critiques by creating new features.

A properly constrained system will make falsifiable predictions about the features that can and cannot be created. We agree about the need for these specific constraints and in fact discussed what some of them might be (see sects. 2.5, 3.3, and 3.5 of Schyns et al. 1998; see also sect. R6 in the same article). For example, Hoffman and Richard's (1984) minima rule (see target article, sect. 2.5) suggests a perceptual bias to create parts with endpoints that are local minima of principle curvature. A flexible theory of part creation that includes this constraint would be falsified if people preferentially created parts with endpoints that are local maxima of principle curvature (see also the accompanying commentaries by **Benson**, **McDorman**, **Sing & Landau**, and **Tanaka** on Schyns et al. 1998, for other specific suggestions). In general, a theory of feature creation will be as good as its constraining principles. These, as argued in the target article, arise from the constraints of perceptual mechanisms, but also from the functional requirements of segmenting the world into specific categories.

R2. Implementing versus identifying constraints

The identification of relevant constraints is the lion's share of the task of modeling feature creation. However, whereas the *implementation* of constraints is done in one specific piece of hardware, their *identification* remains a functional exercise, a point that **Dawson** does not seem to fully appreciate. As Marr (1982, p. 27) famously pointed out, "in order to understand bird flight, we have to understand aerodynamics; only then do the structure of the feathers and the different shapes of birds' wings make sense." Just as the function of flying has been independently converged upon by different organisms using different physical solutions, any number of physical structures may allow a functional constraint to be implemented. Thus, while every functional constraint must be instantiated by *some* physical structure, a physical analysis alone will not reveal the specifications that constrain the development of functional features.

R3. Reduction of functional constraints to physical mechanisms

Dawson argues that functional features might suffer from the scope slip because we view object representations as *(representation of object X) with feature F* instead of the proper *representation of (object X with feature F).* Dawson argues that features should be left outside the functional architecture of the cognitive system under study because, in analogy with neural network models, features do not need to be explicitly represented in connection weights for systems to *behave as* feature detectors.

Whereas we generally agree that networks and their dynamics offer at the moment the better *implementation* of feature creation (see sect. 3.4 of the target article[1]), we believe that the functional level of analysis is mandatory, and also that our proposal does not confuse object properties with their representations (the scope slip). True to our functional analysis of features, in R3.4 we stated that "many . . . information packet[s] can qualify as features as long as the system's psychological response to the packet reveals that it is a discrete, holistic entity in psychological processing." In section R3.3, we discussed a dynamical system that flexibly extracts silhouettes without explicit silhouette detectors. At a functional level, this system satisfies the global computational constraint of extracting long and smooth contours in the image. The system is properly constrained because it predicts that a specific *class* of silhouettes (those made of long and smooth contours) will be preferentially extracted. As required to avoid the scope slip, the system does not confuse object properties with their representation: The network will *behave as* a silhouette detector without an explicit representation of long and smooth silhouettes. The dynamics of the network implements the global constraint with local adjustments of connectivity between edge detectors at a fine granularity. The global constraint of detecting smooth silhouettes is lost in the reduction to the operation of the local network elements.

In sum, constraints do matter for feature creation. Their identification belongs to the computational level of analysis of mechanisms, not the level of their implementation. Properly specified computational constraints can be reduced to the dynamics of local computations, but the implementation of constraints does not itself diminish the need for their functional identification.

NOTE

[1] In fact, several specific proposals of feature creation are statistical principles implemented with networks (see sects. R5 and R6.2).

References

Letters "a" and "r" appearing before authors' initials refer to target article and response, respectively.

Berkeley, I. S. N., Dawson, M. R. W., Medler, D. A., Schopflocher, D. P. & Hornsby, L. (1995) Density plots of hidden value unit activations reveal interpretable bands. *Connection Science* 7:167–86. [MRWD]

Cummins, R. (1983) *The nature of psychological explanation.* MIT Press. [MRWD]

Dawson, M. R. W., Medler, D. A. & Berkeley, I. S. N. (1997) PDP networks can provide models that are not mere implementations of classical theories. *Philosophical Psychology* 10:25–40. [MRWD]

Dawson, M. R. W. & Schopflocher, D. P. (1992) Autonomous processing in PDP networks. *Philosphical Psychology* 5:199–219. [MRWD]

Edelman, G. M. (1992) *Bright air, brilliant fire.* Basic Books. [MRWD]

Marr, D. (1982) *Vision.* W. H. Freeman. [rPGS]

Minsky, M. & Papert, S. (1969/1988) *Perceptrons,* third edition. MIT Press. [MRWD]

Pylyshyn, Z. W. (1981) The imagery debate: Analogue media versus tacit knowledge. *Psychological Review* 88:16–45. [MRWD]

Pylyshyn, Z. W. (1984) *Computation and cognition.* MIT Press. [MRWD]