

following explanation of the US rejection of the Kyoto treaty:

The US Senate rejected Kyoto by a unanimous vote because it failed to include restrictions on developing nations. An agreement that imposes restrictions on developed nations but allows poorer countries to experience all of the gain and none of the pain has just the kind of imbalance that our evolved mechanism of reciprocal altruism forces us to balk at (p. 338).

The evolutionary explanation just seems out of place here. Presumably, the 'evolved mechanism of reciprocal altruism' is something we all inherited from our ancestors. But not everyone balked at the alleged imbalance. Hypotheses about evolved psychological mechanisms cannot even begin to explain why some developed countries ratified the treaty but the US did not. For this we need political science, not evolutionary biology.

After reading *Kindness in a Cruel World*, one might conclude that all the mystery has finally been removed from human fitness altruism, so that it is no longer much of an anomaly for the Darwinian paradigm. But what

about people who act in ways that may be detrimental to their own reproductive fitness, and who do so for the benefit of *future* generations? What sorts of evolved mechanisms could cause people to do *that*? Fitness altruism across the generations still poses a challenge to evolutionary theory.

#### References

- 1 Hamilton, W.D. (1964) The genetical evolution of social behavior. *J. Theor. Biol.* 7, 1–52
- 2 Trivers, R.L. (1971) The evolution of reciprocal altruism. *Q. Rev. Biol.* 46, 35–57
- 3 Axelrod, R. and Hamilton, W.D. (1981) The evolution of cooperation. *Science* 211, 1390–1396
- 4 Thompson, P., ed. (1995) *Issues in Evolutionary Ethics*, SUNY Press
- 5 Barkow, J., Cosmides, L. and Tooby, J. (1992) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, Oxford University Press
- 6 Sober, E. and Wilson, D.S. (1998) *Unto Others: The Evolution and Psychology of Unselfish Behavior*, Harvard University Press

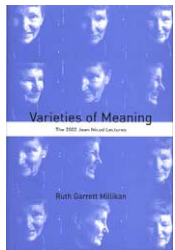
1364-6613/\$ - see front matter © 2005 Elsevier Ltd. All rights reserved.  
doi:10.1016/j.tics.2005.05.007

## The evolution of evolvability

**Varieties of Meaning: The 2002 Jean Nicod Lectures** by Ruth Garret Millikan. MIT Press, 2004. \$35.00/£22.95 (hbk) (242 pages) ISBN 0 262 13444 6

### Thomas Metzinger

Department of Philosophy, Johannes Gutenberg-Universität Mainz, 55099 Mainz, Germany



Ever since Ruth Garrett Millikan burst on the scene with her famous 1984 book *Language, Thought, and Other Biological Categories* [1] she has continued to make substantial contributions, in a remarkably sustained effort that significantly shaped the theoretical landscape in a number of fast-moving fields, from cognitive science to the philosophies of mind, language and biology [1–3]. One of her many achievements lies in the development of a new theoretical approach to cognitive semantics, which philosophers know under the heading of 'teleofunctionalism'. Very roughly, teleofunctionalists share the intuition that, in order to understand the emergence of meaning-carrying states and the instantiation of intentional properties in cognitive systems, we must look beyond computational roles or functional architectures in terms of classical Turing-machine functionalism à la Hilary Putnam [4]. No elaborated, connectionism-inspired microfunctionalism combined with a Churchland-style [5] state-space semantics will do, and even Andy Clark's dynamicist 'epistemic cocoon' [6] made out of transiently incorporated external

memory will not take us all the way. Biological function is what counts in the end. If we want to understand the intentionality of the mental, we have to look at real biological organisms and real evolutionary histories. In her new book *Varieties of Meaning*, which is based on the Jean Nicod Lectures given by Millikan in Paris 2002, she takes a fresh look at the evolution of meaning. Because this book presents an accessible cross-section of Millikan's thinking together with some provocative new ideas in a series of short, crisp chapters, it is bound to attract a considerable number of new readers to her work.

*Varieties of Meaning* is remarkably rich in original ideas, and it offers a wide range of new conceptual instruments. It is hard for me to decide on my favourite ones. The first likely candidate is the distinction between 'ego-implicit' and 'ego-explicit' inner representations. Once the problem of recognizing and reliably re-identifying an 'affording' state of affairs (aspects of the environment that afford the possibility of various activities) under a wide variety of conditions has been solved by an animal, it has to overcome a second obstacle: It must perceive its own *relation* to the affording situation or object in order also to perceive how to perform the necessary manoeuvres, like withdrawing or picking it up or eating it or climbing on it. Millikan writes that both perceptual aspects have to

Corresponding author: Metzinger, T. (metzinger@uni-mainz.de).  
Available online 13 June 2005

be combined in a 'single articulate pushmi-pullyu representation' (p. 175). I of course like this idea, because it nicely fits with some of my own ideas about the internal representation of the intentionality-relation itself [7,8]. However, Millikan goes further by differentiating inner signs, which represent 'enabling relations' by portraying the perceiving animal itself only implicitly, from those which can represent the organism as a whole, and in an explicit fashion. The first category typically models relations of situations and objects to the animal, which are needed for the immediate guidance of its bodily motions. The second type of representations are not immediately involved in action-generation. She interestingly relates this conceptual distinction to neurobiological work on the ventral and dorsal system by Marc Jeannerod [9], criticizing the idea that spatial representation in the brain necessarily needs 'egocentric' or 'allocentric' coordinate systems at all. She then goes on to point out how an explicit or 'objective' representation of the self permits significant transformations in terms of representing entities *other* than the animal in place of the animal's self, thereby allowing it for the first time to be represented as an object among other objects. On the other hand, Millikan writes: 'It may well be that most animals do not have the capacity to represent themselves as objects, so that they harbour only ego-implicit representations and egoless representations, never ego-explicit representations.' (p. 179).

Equally interesting are her ideas of how representations of goal states could become detached. How, in terms of cognitive evolution, can we understand the step from being able to anticipate the consequences of one's own behaviour to projecting a goal, and thereby directly and immediately controlling and guiding one's own activity? The first thing needed is a representation of a goal state, such as a successfully terminated action, enabling the animal to understand if that goal has been reached. Second, the *format* of this representation should be compatible with the system's perceptual or 'descriptive' representations. The general idea is that detached and 'objective' (i.e. neither egoless nor ego-implicit) representations of future, as well as present, states must possess the potential to be matched with projected goal states in a common representational medium, which is functionally unified by a specific format or coding principle. Millikan here touches base with an already existing research program in psychology, the 'common coding' approach [10,11]. Here, as well as in a number of other passages in the book, many readers might have wanted to know if and how all these conceptual distinctions systematically map onto the difference between conscious and unconscious forms of mental content. In detaching and actively generating representations of objects or goal states, are there processing stages which – from a purely teleofunctionalist perspective on mental content – had better be conscious?

The book closes with reflections on the limitations on non-human thought and some speculative ideas about what makes humans special. Millikan says that we might be special in possessing what Dennett [12] calls 'Popperian minds', ones that unfold their intelligence through mentally simulated trials and errors, through continuously 'generating and testing' in multiple inner 'dry runs'. Millikan interestingly points out how free inferential interaction is a central constraint on the medium in which such activity could take place. For distinctively cognitive systems, a mental format is needed that exhibits an articulation into subject and predicate, and that is sensitive to internal negative transformation. Millikan calls this the development of theoretical concepts and theoretical knowledge (p. 216ff). Such representations would not be dependent on their satisfaction conditions, and a representational system adapted to the need of safely guiding an animal's continuous motions through its immediate environment could hardly achieve this, Millikan supposes. However, and this may be the centrally important thought at the end of this innovative and stimulating book, human beings clearly seem to be special in that they internally represent time, not as a mere set of conditional probabilities with regard to temporal relations, but as explicitly *dated* or 'historical' time. In Millikan's own words: 'An animal that represents time as it represents space moves into the future as if navigating a terrain that is already there [...]. We humans who represent time as historical understand that we are *constructing* a sequence, not finding one.' (p. 228).

## References

- 1 Millikan, R.G. (1984) *Language, Thought, and Other Biological Categories*, MIT Press
- 2 Millikan, R.G. (1993) *White Queen Psychology and Other Essays for Alice*, MIT Press
- 3 Millikan, R.G. (2000) *On Clear and Confused Ideas*, MIT Press
- 4 Putnam, H. (1967) Psychological predicates. In *Art, Mind, and Religion* (Capitan, W.H. and Merrill, D.D., eds). Reprinted 1975 as Putnam, H. 'The nature of mental states'. In *Mind, Language, and Reality: Philosophical Papers Vol. 2*, MIT Press
- 5 Churchland, P.M. (1986) Some reductive strategies in cognitive neurobiology. *Mind* 95, 279–309
- 6 Clark, A. (2003) *Natural-Born Cyborgs: Minds, Technologies and the Future of Human Intelligence*, Oxford University Press
- 7 Metzinger, T. (2003) *Being No One*, MIT Press
- 8 Metzinger, T. (in press) Conscious volition and mental representation: Towards a more fine-grained analysis. In *Disorders of Volition* (Sebanz, N. and Prinz, W., eds), MIT Press
- 9 Jeannerod, M. (1997) *The Cognitive Neuroscience of Action*, Blackwell
- 10 Prinz, W. and Hommel, B., eds (2002) *Common Mechanisms in Perception and Action. Attention and Performance (Vol. XIX)*, Oxford University Press
- 11 Hommel, B. *et al.* (2001) The theory of event coding (TEC): a framework for perception and action planning. *Behav. Brain Sci.* 24, 848–878
- 12 Dennett, D.C. (1996) *Kinds of Minds*, Harper Collins