Towards a philosophy for neuroethics

An informed materialist view of the brain might help to develop theoretical frameworks for applied neuroethics

Kathinka Evers

Early in the twenty-first century, neuroscience has developed into a major research field that is rapidly spawning new academic disciplines. One of these is neuroethics, which has been described as an “unexplored continent lying between the two populated shores of ethics and of neuroscience” and “a new area of intellectual and social discourse” (Roskies, 2002). However, neuroethics is not a new research area: for a long time, ethicists have dealt with problems arising from advances in neuroscience, although not necessarily under the label of neuroethics.

So far, neuroethicists have focused mainly on applied neuroethics, such as the ethical issues associated with neuroimaging techniques, cognitive enhancement and neuropharmacology (Illes, 2006). A less prevalent focus is fundamental neuroethics, which investigates how knowledge of the functional architecture of the brain and its evolution can increase our understanding of personal identity, consciousness and intentionality, including the development of moral thought and judgement. Fundamental neuroethics therefore provides applied neuroethics with the theoretical foundations needed to address ethical problems of applying neurological science.

The relevance of neuroscience to fundamental neuroethics depends on which theoretical model of the brain is used to understand complex human behaviour. A naive reductionist model—which postulates that, as consciousness is the product of neuronal activity in the brain, we only have to wait for neuroscience to become sufficiently sophisticated to be able to explain all that there is to know about it (Crick, 1994; Bickle, 2003)—is no longer acceptable.

Neuroethics articles continue to speak of “the astonishing hypothesis” that “the mind is the brain” (Illes & Racine, 2005), but this is either trivial or absurd. The insight that mind and brain are not dual entities might have been astonishing when it was first realized, but it is now common scientific knowledge. However, this does not imply that objective knowledge about neuronal activities in the brain can tell us all that there is to know about consciousness without referring to subjective accounts that are based on self-reflection. No objective third-person description can fully capture the first-person view from its own perspective; it needs to be experienced first-hand to be known, and to live the experience of another is a logical impossibility. Fundamental neuroethics must therefore acknowledge both the embodiment of consciousness and the irreducible subjective perspective. Moreover, as the subjective experience of the brain mobilizes emotions that constitute an essential feature of moral judgement, fundamental neuroethics must also take into account the role of emotions and values.

Here, I suggest a model of the brain that might provide a theoretical framework for neuroethics: informed materialism. It has a high explanatory value with regard to the evolution and nature of moral judgement, notably under the aspects of subjective evaluation, free will and personal or public responsibility. The concept of informed materialism was originally used in chemistry (Bachelard, 1953), but has been used in neuroscience to oppose both dualism and naive reductionism (Changeux, 2004). It is founded in the notion that all elementary processes of the brain are based on physicochemical

mechanisms, and adopts an evolutionary view of consciousness as a biological function of neuronal activities. However, this model also describes the brain as an autonomously active, projective and variable system, in which emotions and values are incorporated as necessary constraints. Owing to the way in which our brains acquire knowledge of ourselves and the world, informed materialism acknowledges that adequate understanding of our subjective experience must take into account self-reflective information, physiological observations and physical measurements (Dehaene *et al*, 2003; Changeux, 2004).

The concept of informed materialism [...] has been used in neuroscience to oppose both dualism and naive reductionism

In this model, emotions have an important role; some researchers propose that they mark the dawn of consciousness (Denton, 2005), motivation, planning and volition, and increase the capacity of an organism to exert control over itself (LeDoux, 1996). Current hypotheses of knowledge acquisition posit that spontaneously arising pre-presentations are selected by reward signals, and are stabilized as representations after confirmation by both external experience and internal evaluation processes (Dehaene & Changeux, 1991). Such models of the world are stabilized through cognitive games as permanent features of a developing cognitive apparatus (Dehaene *et al*, 1998; Changeux, 2004). The anticipation of a reward creates a delay between the elaboration of tacit plans and the actual interaction

with the world, which is an additional component of cognitive learning and a building block for the neural basis of morality (Schultz, 2006; Schultz *et al.*, 1997).

A non-emotional system, by contrast, would be passive and incapable of self-organization, evaluation or anticipation, and hence learning; without any values, a system cannot learn or remember (Edelman, 1992). The complex behaviour of humans might not have evolved without the basic feature of consciousness, and consciousness could not have evolved without emotions. From the perspective of an informed materialist, dynamic processes of evaluation and the emotional systems that are involved in their generation are basic properties of our brains: we are neurobiologically predisposed to develop complex and diverse systems of moral and other values that allow us to establish appropriate relationships in our social, cultural and physical environments.

The notion of free will is a fundamental moral concept that is central to human experience. Having the freedom to choose is, as a rule, regarded as a necessary condition for moral agency and personal responsibility, around which all societies are built. With reference to neuroscience, our experience of having free will has been considered illusory by virtue of being a useful but fictional brain construct (Smith Churchland, 2002) that is causally determined (Blakemore, 1988) or non-consciously initiated (Libet, 1999). According to the model suggested here, neither view is valid: an act of will can be 'free' in the sense of being 'voluntary', even if it is a brain construct that is causally determined or influenced by non-conscious neuronal processes.

The brain might generate illusions under certain conditions, but the fact that an experience is a brain construct does not make it illusory. The ability to distinguish between the illusory and the real is necessary for the survival of the organism, and neuronal causation cannot be a token of the illusory, lest all we experience is an illusion.

The complex behaviour of humans might not have evolved without the basic feature of consciousness, and consciousness could not have evolved without emotions

To the extent that our choices are not causally determined or, on the contrary, causally predetermined, our experience of voluntary decision-making might seem illusory: in the first case, we are generators of chance; in the second, we live on autopilot. Informed materialism rejects both theories. The brain is not a causal system that functions with conditional necessity; it admits random events and variable outcomes. This contradicts the naive picture of the brain as a rigid automaton the operation of which is strictly predetermined (Blakemore, 1988)—a model that would have low explanatory value with respect to morality, as automatons are not moral agents.

The brain is, consciously and non-consciously, a volitional organ (Libet *et al.*, 1999; Ingvar, 1999), which is distinguished

by its plasticity, and by the general capacity of the neuron and its synapses to change properties as a function of their state of activity. The brain is a gigantic network of neurons, the distribution of which is largely limited by genetics, but is simultaneously dependent on random variations that allow the network to react to physical, social and cultural inputs (Changeux, 1985). This variability implies a form of plastic determinism in which causal antecedents can produce variable outcomes.

The variability theorem states that distinct learning inputs can produce varying connective organizations and neuronal abilities, but similar behavioural abilities (Changeux *et al.*, 1973; Changeux, 2004). This is a consequence of the 'synapse selection' in which new combinatory patterns

are continually produced and tested as a neural embodiment of creative activity. This happens whenever the brain produces communicable representations of the world despite phenotypic variance of its synaptic connectivity. The ability of epigenesis—by selectively stabilizing neurons and synapses during development—to take variability into account is one of its main outcomes (Changeux, 1985). An additional advantage is that free will reappears; plastic determinism restores an element of voluntary influence that can act as a basis for personal responsibility.

The brain might generate illusions under certain conditions, but the fact that an experience is a brain construct does not make it illusory

The nature of this influence might be largely non-conscious. From a neurophysiological point of view, consciousness is only the tip of a mental iceberg (LeDoux, 1996; Dehaene *et al.*, 2006), which, by virtue of its limited capacity, continuously needs to liberate space, such as by developing routine non-conscious behaviour. It has been argued that we have no control over, and therefore are not responsible for, non-consciously motivated acts, which might be volitional but not voluntary and hence not free (Libet, 1999). These assumptions are questionable. Admittedly, we cannot be held responsible for an act that is not the result of conscious deliberation; however, a distinction between voluntary and non-voluntary can be drawn in both the conscious and the non-conscious mental realms. Underlying our conscious experiences are brain mechanisms of which we are unaware—notably, non-conscious emotions, cognitive programmes and stored memories—but our conscious neuronal workspace and processes are not causally disconnected from these mechanisms. The causal influence between the conscious and the non-conscious is not unilateral but instead mutual, although asymmetrical. The possibility of consciously influencing non-conscious mental mechanisms is a crucial assumption in any programme of education, self-improvement or therapy, and must be taken into account to determine the role of consciousness in producing actions. If their influence is mutual,

non-conscious mechanisms might pose a challenge to, and sometimes reduce, conscious control without excluding it in principle.

Under the model of informed materialism, it is not reasonable to maintain that only conscious decisions can properly be described as free. The brain, in its social and cultural context, exhibits will through the capacity to desire, and free will through the capacity to strive for the satisfaction of this desire. Free will would therefore be the capacity to acquire causal power coupled with the capacity to influence the use of that power. This power creates responsibility. To the extent that we might influence our non-conscious, we can be personally responsible for behaviour resulting from these non-conscious influences. Given a certain level of maturity, the brain—functioning in a given social and cultural environment—can accordingly be regarded as a ‘responsible organ’. In this model, humans act as free and responsible agents, while being contingently causally determined and influenced by non-conscious processes that are not beyond conscious control. Rather than posing a new threat to the morally fundamental notions of free will and personal responsibility, neuroscience can offer them empirical support.

This implies that evolution and natural selection have created an evaluative cerebral architecture with a propensity for performing free and responsible moral choices, and that morality is a biologically conditioned social phenomenon that has evolved genetically and epigenetically through the communication and interaction of self-aware organisms. Self-awareness—the capacity to identify oneself as a subject of experience—is typically accompanied by self-interest, which is an innate preference that creates an urge to control the immediate environment.

In social creatures, self-awareness also inspires interest in others, primarily those to whom the self can relate and with whom it identifies, such as the next of kin (Singer *et al.*, 2006). In intelligent social species, the interest in the self—already active in the newborn—is extended to endorse close family members and then the social group, but distinctions are soon drawn between ‘we’ and ‘they’ (Decety, 2007). We typically extend sympathy and aid to others depending on their closeness to us in terms of innate dispositions

(Hart *et al.*, 2000), but also in relation to epigenetically acquired differences in culture, ideology and so forth. This is where the field of ethics arises. In distinguishing the self from others, and drawing lines between ‘we’ and ‘they’, a hierarchy becomes established at the social level, and moral theory becomes a wealth of analyses and proposals regarding the relationships between these different levels (Ricoeur, 1992). Conflicts within a society or community, or between individuals, are related mostly to group distinctions and to the distribution of privileges (Changeux & Ricoeur, 2000).

...morality is a biologically conditioned social phenomenon that has evolved genetically and epigenetically through the communication and interaction of self-aware organisms

The genetic control over the development of the brain is therefore not absolute: the epigenetic model of neuronal development postulates that the connections between neurons are not pre-specified in the genes, but instead that learning and experience influence the development of the brain within the boundaries of a ‘genetic envelope’.

The formation of synapses takes place both prenatally and postnatally; it is far from complete at birth, and the postnatal development of the human brain lasts considerably longer than in any other mammalian species (Bourgeois, 1997). The most intense development occurs during the first two years, but the process continues long after puberty; in fact, the highest executive functions determined by the frontal lobe are not fully mature until the age of around 20 years. The environment is important for this process. If neural networks do not stay active they vanish, possibly irreversibly (Katz & Shatz, 1996). The socio-cultural structures of the group and the neuronal functions of the individual therefore develop symbiotically. The architecture of our brains contributes to our social behaviour and the types of society that we create, which in turn influence the development of the brain. The neuronal features that develop as a result of learning and experience can therefore be passed on through generations.

This normative relevance of neuroscience should not be understood in the naive sense of describing facts that entail or constitute norms. It is commonly accepted that normative judgements, moral or otherwise, cannot logically be derived from descriptive statements—committing what in philosophy is known as “the naturalistic fallacy” (Moore, 1903). This fallacy has different versions; the one that is relevant here consists of deriving an ‘ought to be’ from an ‘is’, which is a conceptual mistake that informed materialism does not make.

Fundamental neuroethics must [...] acknowledge both the embodiment of consciousness and the irreducible subjective perspective

Mere factual descriptions of the functional architecture of the brain are not tantamount to making recommendations or assertions of norms; nevertheless, there are important empirical, notably causal, connections between biological facts, norms and moral values. Norms are brain constructs that are biologically, as well as culturally, embedded in and constrained by socio-cultural structures. A major responsibility of fundamental neuroethics is to decipher this network of causal connections between the neurobiological, socio-cultural and contingent historical perspectives that gave rise to moral norms at a given moment in human history, and to evaluate their ‘universal’ character as pre-specified in our genome and shared by members of the human species. The ‘fallacy’ of the naturalistic approach therefore becomes inverted into a responsibility.

Neuroscientific knowledge can deepen our understanding of who we are, and of how we function as neurobiological and social creatures. It can help to explain the mechanisms of normative judgement and how morality has evolved. This knowledge might allow us to solve social problems, improve our mental, physical and social health, refine our education systems and develop our societies (Evers, 2005). By contrast, it can also be put to civilian or military misuse, and neuroethics must maintain a high level of vigilance in this regard (Evers, 2007). As scientific responsibility cannot arise in the absence of scientific adequacy,

applied neuroethics needs plausible theoretical frameworks developed by fundamental neuroethics. An informed materialist view of the brain might provide a scientifically adequate and philosophically fruitful start to develop such frameworks and a philosophy for neuroethics.

ACKNOWLEDGEMENTS

I thank J.-P. Changeux for his contributions to this article and his helpful comments.

REFERENCES

- Bachelard G (1953) *Le Matérialisme Rationnel*. Paris, France: Presses Universitaires de France
- Bickle J (2003) *Philosophy and Neuroscience. A Ruthlessly Reductive Account*. Dordrecht, The Netherlands: Kluwer
- Blakemore C (1988) *The Mind Machine*. London, UK: BBC Publications
- Bourgeois JP (1997) Synaptogenesis, heterochrony and epigenesis in the mammalian neocortex. *Acta Paediatr Suppl* **422**: 27–33
- Changeux JP (1985) *Neuronal Man*. New York, NY, USA: Pantheon Books
- Changeux JP (2004) *The Physiology of Truth*. Cambridge, MA, USA: Belknap
- Changeux JP, Ricoeur P (2000) *What Makes us Think? A Neuroscientist and a Philosopher Argue About Ethics, Human Nature, and the Brain*. Princeton, NJ, USA: Princeton University Press
- Changeux JP, Courrègne P, Danchin A (1973) A theory of the epigenesis of neuronal networks by selective stabilization of synapses. *Proc Natl Acad Sci USA* **70**: 2974–2978
- Crick F (1994) *The Astonishing Hypothesis: The Scientific Search for the Soul*. New York, NY, USA: Scribner
- Decety J (2007) A social cognitive neuroscience model of human empathy. In Harmon-Jones E, Winkelman P (eds) *Social Neuroscience: Integrating Biological and Psychological Explanations of Social Behavior*, pp 246–270. New York, NY, USA: Guilford
- Dehaene S, Changeux JP (1991) The Wisconsin card sorting test: theoretical analysis and modeling in a neuronal network. *Cereb Cortex* **1**: 62–79
- Dehaene S, Kerszberg M, Changeux JP (1998) A neuronal model of a global workspace in effortful cognitive tasks. *Proc Natl Acad Sci USA* **95**: 14529–14534
- Dehaene S, Sergent C, Changeux JP (2003) A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proc Natl Acad Sci USA* **100**: 8520–8525
- Dehaene S, Changeux JP, Naccache L, Sackur J, Sergent C (2006) Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends Cogn Sci* **10**: 204–211
- Denton DA (2005) *Les Émotions Primordiales et l'Éveil de la Conscience*. Paris, France: Éditions Flammarion
- Edelman GM (1992) *Bright Air, Brilliant Fire: On the Matter of the Mind*. New York, NY, USA: BasicBooks
- Evers K (2005) Neuroethics: a philosophical challenge. *Am J Bioeth* **5**: 31–33
- Evers K (2007) Perspectives on memory manipulation: using beta-blockers to cure post-traumatic stress disorder. *Camb Q Healthc Ethics* **16**: 138–146
- Hart AJ, Whalen PJ, Shin LM, McInerney SC, Fischer H, Rauch SL (2000) Differential response in the human amygdala to racial outgroup vs ingroup face stimuli. *Neuroreport* **11**: 2351–2355
- Illes J (ed; 2006) *Neuroethics: Defining the Issues in Theory, Practice, and Policy*. Oxford, UK: Oxford University Press
- Illes J, Racine E (2005) Imaging or imagining? A neuroethics challenge informed by genetics. *Am J Bioeth* **5**: 5–18
- Ingvar DH (1999) On volition: a neurophysiologically oriented essay. In Libet B, Freeman A, Sutherland K (eds) *The Volitional Brain: Towards a Neuroscience of Free Will*, pp 1–10. Thorverton, UK: Imprint Academic
- Katz LC, Shatz CJ (1996) Synaptic activity and the construction of cortical circuits. *Science* **274**: 1133–1138
- LeDoux J (1996) *The Emotional Brain*. New York, NY, USA: Simon & Schuster
- Libet B (1999) Do we have free will? In Libet B, Freeman A, Sutherland K (eds) *The Volitional Brain: Towards a Neuroscience of Free Will*, pp 47–58. Thorverton, UK: Imprint Academic
- Libet B, Freeman A, Sutherland K (eds; 1999) *The Volitional Brain: Towards a Neuroscience of Free Will*, pp 47–58. Thorverton, UK: Imprint Academic
- Moore GE (1903) *Principia Ethica*. Cambridge, UK: Cambridge University Press
- Ricoeur P (1992) *Oneself as Another*. Chicago, IL, USA: University of Chicago Press
- Roskies A (2002) Neuroethics for the new millennium. *Neuron* **35**: 21–23
- Schultz W (2006) Behavioral theories and the neurophysiology of reward. *Annu Rev Psychol* **57**: 87–115
- Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science* **275**: 1593–1599
- Singer T, Seymour B, O'Doherty JP, Stephan KE, Dolan RJ, Frith CD (2006) Empathic neural responses are modulated by the perceived fairness of others. *Nature* **439**: 466–469
- Smith Churchland P (2002) *Brain-Wise: Studies in Neurophilosophy*. Cambridge, MA, USA: MIT Press

Kathinka Evers is Associate Professor of Philosophy and Senior Researcher at the Centre for Bioethics at Karolinska Institutet and Uppsala University in Uppsala, Sweden.

E-mail: kathinka.evers@bioethics.uu.se

doi:10.1038/sj.embor.7401014