Neuroethics, a New Frontier in the Humanities

Stanislas DEHAENE

What does the science of the brain tell us about our moral judgments and behavior? Two books address that question in presenting a new discipline: neuroethics.

Reviewed: Bernard Baertschi, *La neuroéthique: ce que les neurosciences font à nos conceptions morales*. Éditions La découverte, 2009; Kathinka Evers, *Neuroéthique: quand la matière s'éveille*. Editions Odile Jacob, 2009.

Chemists and biologists know full well that the most active reactions and the most radical developments invariably occur at the frontier where two hitherto separate fields suddenly brush up against each other. Ideas emerge according to the same law, as it is often on the fringes of two disciplines that fundamentally novel propositions arise. In contemporary science, one of the fastest-moving frontiers is that of cognitive neuroscience, the branch of neuroscience that seeks to understand the neuronal mechanisms of human capabilities by combining methods of psychology and brain imaging. Self-awareness, moral judgment, the impact of education and plenty of other questions formerly subsumed under moral philosophy have become new subjects of research for neuroscience. This seemingly expansionist hyperactivity has given rise to a legitimate fear: Won't the materialist acid of neuroscience corrode or even destroy the very pillars of our society: free will, responsibility, individual identity and moral judgment?

These questions are addressed by a nascent discipline commonly called "neuroethics," at the interface between philosophy and neuroscience. Two excellent recent books in French present an in-depth view of this new field of inquiry. Bernard Baertschi, who teaches and researches at the Institute of Biomedical Ethics and in the Philosophy department in Geneva, does most of his work in classical philosophy. As the title of his book, *La neuroéthique : ce que les neurosciences font à nos conceptions morales* ("Neuroethics: What Neuroscience Does to Our Moral Conceptions"), suggests, the object is mainly to put the leading philosophical theories of moral judgment, responsibility and personal identity into perspective in the light of recent discoveries in neuroscience. As he puts it to his fellow philosophers, "If it was already wise back in the 18th century to take what science had to teach us into consideration when addressing philosophical issues, it has become absolutely imperative in our day and age."

Katinka Evers, a professor at the Center for Research in Ethics and Bioethics at the University of Uppsala (Sweden), while basing her arguments on a very similar empirical and philosophical approach, sets herself an even more ambitious goal. Her book, *Neuroéthique :*

quand la matière s'éveille ("Neuroethics: When Matter Awakens"), the upshot of a lecture series given at the Collège de France, outlines a novel philosophical view of the human condition, an "enlightened materialism" that seeks to reconcile our undeniably neurobiological nature with a humanist approach attentive to individuals and their harmonious lives in society. Building on Jean-Pierre Changeux's arguments in *Neuronal Man: The Biology of Mind* [found Eng. translation according to wiki] (1985) and *The Physiology of Truth* (2004), two books that can legitimately be called pioneering works in neuroethics, she aptly observes, "We are neuronal men and women in the sense that everything we do, think and feel is a function of the architecture of our brains; and yet this fact has yet to be fully incorporated into our general conceptions of the world and of ourselves" (p. 32). The age-old dualism that places mental functions on a separate plane from those of the body and consequently of the brain still pervades every domain of society: our churches and religions, of course, but also our legal system, our social organization and our schools (will our ideas about merit and talent withstand advances in neuroscience?). Both books are demanding, sometimes downright difficult, but always highly stimulating, and they herald an imminent upheaval in received ideas.

What is neuroethics?

Neuroethics actually encompasses at least two domains, which Baertschi calls "the ethics of neuroscience" and "the neuroscience of ethics," while Evers makes a similar distinction between basic and applied neuroethics:

> Basic neuroethics looks into the ways in which knowledge of the functional architecture of the brain and its evolution can deepen our understanding of personal identity, conscience and intentionality, including the development of moral thought and judgment. Applied neuroethics studies the ethics of neuroscience, for example the ethical problems raised by neuron imaging technologies, cognitive improvement and neuropharmacology." (K. Evers, p. 204-205)

The two books also agree on the main issues to be addressed by neuroethics: (1) How to conceive of responsibility and free will, particularly from a legal viewpoint, if man is but a neuronal machine? (2) On what foundation can we base our ethical principles, if our moral sense is merely the product of a brain tinkered together by evolution, in which oft-conflicting mechanisms of emotional and rational judgment coexist? (3) How to revise social norms with a view to providing a "good life" to everyone, and what limits to impose on neurotechnological development, given that neuroscience will soon make it possible to alter the workings of the brain, for the better or for the worse?

Faced with these somewhat frightening prospects, our two authors for the most part show a great deal of reserve – as well as an optimism which I share and which forms a congenial contrast to the paralyzing virtual omnipresence of the precautionary principle. "Ethics has nothing to fear from neuroscience, so there's no cause for moral panic!" Braetschi reassures us. "Meaning and meaningfulness, our dignity and *raison d'être*, and other concepts so dear to humankind are not lost to us simply because they happen to be constructs that depend on our cerebral architecture," concurs Evers.

To sum up why that is so would be a formidable task, as these two books are densely packed with arguments. Instead, I will attempt to bring out a few key points, putting my own two cents in where appropriate. First of all, neuroscience has lost a bit of the arrogance that once induced it to contemplate casting out the whole vocabulary of psychology from the field

of natural science. Conversely, cognitive psychology is no longer leery of taking an interest in the architecture of the brain or exploring subjects hitherto deemed taboo and impossible to analyze scientifically, such as consciousness, self-awareness and emotions. The ultra-reductionism of the behaviorists has had its day, as has the naïveté of cognitivism based entirely on the metaphor of the computer. They have given way to cognitive neuroscience, a "psychophilic science," as Evers puts it, that is more homogeneous and concomitantly more multidisciplinary, that no longer writes off mental or emotional states as immaterial fictions diametrically opposed to physical reality, but on the contrary seeks to shed some light on the neuronal architectures underlying those states of mind.

Let's take consciousness, for example, a subject of extensive empirical research in our day. Researchers no longer deny that taking cognizance of a piece of information involves a significant change in the state of activity of our neuronal networks. To identify the mechanisms behind that change, researchers need to simultaneously obtain subjective psychological information (verbal reporting by subjects describing their perceptions) and objective neuronal data (cerebral activity measured by electroencephalography or magnetic resonance imaging). According to this materialist but non-reductionist view of consciousness, the human brain is a biological organ endowed by Darwinian evolution with a structured architecture that projects mental representations, assessments, a personal perspective (a "self") and plans of action onto the surrounding world. This neuronal organization is variable and flexible: it is not wholly subjugated to the dictates of our genes, but capable of incorporating cultural conventions and moral rules. And consciousness is part and parcel of its most basic material properties.

Free will endures

Within the framework of this "enlightened materialism," there is no contradiction in talking about the cerebral foundations of moral responsibility. The concept of free self-willed choice doesn't melt away in the heat of advances in neuroscience. It is a profound error to deny that free will endures under the pretext of the possible determinism of nerve activity, or to look for the foundations thereof in basic quantum uncertainty. Indeed, those who think that our freedom will necessitates a throwing of the quantum dice are making a mistake comparable to that of a physicist attempting to explain the solidity of a house in terms of the atomic structure of its wooden beams. In reality, what we (legitimately!) call free will is but the functional description of a brain which, through its very organization, possesses the capacity to consider several different ways of acting, to assess the consequences thereof and to choose one of them, without that choice being wholly predictable from outside or even inside the subject's brain. Thus, argues Baertschi, "there is no need for free will *sensu stricto* [i.e. a rift in physical causality] in order to bring in responsibility; a lucidly desired intentional action will suffice." Consequently, neuronal man remains a responsible person, at least if his brain has not suffered alterations or lesions.

Neuroscience itself helps to define this normal mode of brain functioning: intentional, capable of reflecting fully consciously and lucidly. Brain imaging has indeed revealed huge differences between actions we carry out automatically, sometimes wholly unconsciously, and those that involve a thought-out choice, those that we control. The prefrontal cortex, also known as the "central administrator," plays a key role in this executive control, regulating,

choosing or inhibiting according to the goals we set ourselves.[1] The prefrontal cortex, the section of the cerebral hemispheres situated right below the forehead, constitutes up to 30% of the surface of the cortex in man, a proportion unrivalled in any other primate. Its smooth functioning is essential to every considered decision. Even if our knowledge of its organization remains fragmentary, it can nonetheless legitimately be called the cerebral seat of reason, of planning and mental synthesis.

The emotional factor

To say that our choices are rational and conscious by no means rules out their being guided, at least in part and sometimes unconsciously, by our emotions. In the new conceptual framework emerging from the latest neuroscientific research, the emotional circuits responsible for our rapid physical reactions (fear, joy, anger etc.) as well as the main routes that liberate neuromodulating chemical agents (e.g. dopamine, noradrenaline, acetylcholine) constitute the building blocks of sophisticated algorithms for the assessment of our past and future situations on different temporal scales. Their mixed signals constitute the pillars of a value system on the basis of which we decide whether or not to modify our behavior at any given moment.[2]

So advances in neuroscience help us to better discern the boundaries within which we make our free and responsible choices. The harmonious functioning of cerebral systems of emotion appears to be essential to rational conduct, and their perturbation reduces the range of choices available to the subject, sometimes to the point of disabling him from exercising any control over his behavior. Addiction, for example, occurs when a pharmaco-chemical agent diverts from their primary purpose the decision-making circuits that rely on dopamine and acetylcholine to constantly adapt our choices as we learn over time according to the reward signals we receive from the environment. Theoretical models of these neuronal circuits[3] explain how drugs, even when they cease to give us pleasure, continue to bias our decision making to the point of making the addict a being not devoid of willpower, but whose every decision is reoriented towards a pathological objective. Though Evers and Baertschi might have elaborated on this example, they opt instead to follow up on the work of neuroscientist Antonio Damasio to explore the well-known case of patients like Phineas Gage who suffer from lesions of the orbitofrontal and medial cortex, which render their bodies and brains insensible to emotional signals associated with the anticipation of consequences, whether good or bad, of intended actions.[4] These people make poor decisions in life. Their conduct, which has become amoral, often leads them to lose their jobs, friends and family, and in some cases lands them in prison, even though they remain capable of verbally articulating the moral rules which their sick brains no longer allow them to follow. Even this latter competence, which might be called a theory of morality, can disappear if the lesion occurs very early in childhood, thereby producing adolescents at the mercy of their own extremely antisocial utilitarian and selfish urges.

---

[1] Passingham, R. (1993). *The Frontal Lobes and Voluntary Action* (Vol. 21). New York: Oxford University Press.
[2] Yu, A. J., & Dayan, P. (2005). "Uncertainty, Neuromodulation, and Attention". *Neuron, 46*(4), 681-692.
[3] Gutkin, B. S., Dehaene, S., & Changeux, J. P. (2006). "A Neurocomputational Hypothesis for Nicotine Addiction". *Proc Natl Acad Sci USA, 103*(4), 1106-1111; Redish, A. D. (2004). "Addiction as a Computational Process Gone Awry". *Science, 306*(5703), 1944-1947.
[4] Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain.* New York: G.P. Putnam.

As we can see, neuroethical thought points to what might be termed "attenuating neuronal circumstances" for certain individuals whose brains don't work normally, though without – and this is the crucial point – calling for the eradication of the concept of individual responsibility. It is simply a matter of accepting, and integrating into society, the fact that human nature radically diverges from the idealized dualistic conception that our current legal system ascribes to it. We are the result of some fairly strange evolutionary tinkering, a mass of circuits that proved beneficial for survival in the past; so there is no guarantee of their being consistent, in fact they never cease to tug in opposite directions.

Multiple cerebral bases of moral judgments

Insofar as a goodly part of our evolution takes place within society, part of our cerebral paraphernalia governs social conduct. Adults, children, and even some other primates all have a heightened sense of justice, which makes us reject an unequal distribution of goods, for instance, even at our own expense – as clearly demonstrated by a test that has become a classic, the so-called "ultimatum game." The altruistic choice to share for the greatest common good, and to extend sympathy beyond our own biological family to the entire social group, could result from a combination of this sense of justice and our peculiar ability to represent others' states of mind – what psychologists call our "theory of mind." However, other neuronal structures militate against these pro-social tendencies. The brain's amygdala, a bilateral mass of gray matter specialized in the rapid assessment of emotional or dangerous situations, reacts in a hypersensitive manner to faces that are unfamiliar or of ill repute. An out-and-out subliminal racism is inscribed in this circuit, whose rapid discharges to a stranger's face engender a state of fright even before we've become aware of the presence of the face. New research by Elizabeth Spelke at Harvard University suggests that in a child's first year of life, it is spoken language, even more than race or skin color, that induces the spontaneous rejection of the other.

As Katinka Evers and Bernard Baertschi point out, this contradictory cast of the human brain, this mix of altruistic and xenophobic circuits, hardly leaves any hope of legitimizing this or that ethical conception on a neuronal basis. As usual, science says what *is*, not what *should be*. It's up to society to construct systems of moral judgment. Nevertheless, those systems will gain in pertinence and in humanity to the extent that they take into account the nuanced portrait of our human condition provided by neuroscience.

And Evers adds a crucial point: individuals are not wholly determined by their genes or their personal histories. An ongoing variability characterizes neuronal activity, which can swing from the heroic to the heinous and vice versa in a manner that will always remain unpredictable. This variability, particularly conspicuous in the neurons of the cerebral cortex, does not disappear over the course of human development, but is modulated by the tremendous malleability of the human brain. Man is the only primate who is born so immature that, up to and even beyond adolescence, millions of synapses continue to be created and eliminated every day. This innate variability and malleability enable us to learn to regulate our behavior based on epigenetic rules laid down by our social group. Upbringing and education can be viewed as processes of channeling neuronal variability in what are deemed to be desirable directions. In my opinion it is here, in the domain of neuro-education, that neuroethical thought becomes most interesting. Through education, ours is the only species capable of choosing its own ethics. In elucidating the neuronal mechanisms behind our moral judgments and in gaining a better understanding of how they are modulated by the education

we have received, we will empower ourselves to exercise greater control over our acts. According to Baertschi, not only would it be wrong for us to shy away from exploring this new power, but from an ethical point of view we will soon be duty-bound to act if and when science makes it possible to modify the brain in a way that would ensure a happy or decent life for those on the slippery slope towards moral trouble owing their cerebral architecture.[5]

Beware of "neuro-myths"!

But to discuss these matters calmly, a solid empirical basis is of the essence. And this may be where our two introductions to neuroethics come up short. The empirical results they describe remain meager at best, too sketchily outlined and often somewhat outdated. Baertschi actually succumbs to a certain sensationalism verging at times on science fiction. Is it really appropriate in our day to talk about the prospect of erasing undesirable memories (p. 14), a "compliferant substance" that will utterly annihilate the will (p. 15), transplanting human brain cells in animal brains (p. 18), the "little protuberance of gray matter" responsible for religious or mystical experiences (p. 98) or the "God helmet" (p. 99) that allows anyone else to share those experiences? The inquisitive reader will search in vain for serious scientific references on these provocative subjects, and find nothing but some tentative research, even some "neuro-myths" spread by certain neuroscientists eager for media recognition, which the author would be better off dissipating rather than propagating.

Katinka Evers' book, far more rigorous on this head, doesn't shy from consigning certain pipedreams to the oblivion where they belong in a single lethal line: "One example of an unscientific suggestion," she says, "is the purported new possibility of 'reading minds' by means of electronic implants in the brain or scanners." As a neuroscientist myself, I can only concur, while regretting that Evers does not more fully explain to her readers why it's unscientific. Actually, the technology has made significant headway, to a point where magnetic resonance imaging, combined with powerful algorithms for image classification, can decode certain mental representations. Recordings of brain activity now make it possible to determine which of two numbers has been shown to a subject[6], for example, or which of two words[7]. These results are profoundly important to basic research insofar as they reveal the existence of a neural code of symbols and words that is distributed to the entire cortex, detectable with millimeter precision, and that is partly shared by every single person. These astonishing findings refute virtually everything we thought we knew about the representation of the meaning of words in the human brain. However, this research doesn't mean we can "read the brain" in a way that might worry neuroethicists. "Brain-reading" experiments typically show a very limited success (about 60%, compared to the 50% chance level expected from a binary choice made purely at random), and carried out with willing and immobilized volunteers – conditions not likely to be produced without the subject's knowledge or against his will. The findings are, however, of great clinical interest. Thanks to this research, patients who are aphasic, hemiplegic or "locked-in" – i.e. suffering from the syndrome described by stroke victim Jean-Dominique Bauby in his memoir *The Diving Bell*

[5] See also Dennett, D. (2003). *Freedom evolves*. London: Allen Lane.

[6] Eger, E., Michel, V., Thirion, B., Amadon, A., Dehaene, S., & Kleinschmidt, A. (2009). "Decoding of individual number information from spatial activation patterns in human intraparietal cortex". *Current Biology*, forthcoming.

[7] Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A. et al. (2008). "Predicting human brain activity associated with the meanings of nouns". *Science, 320*(5880), 1191-1195.

and the Butterfly – will soon learn to communicate again and to control their bodies using brain-machine interfaces.

In this domain, as in plenty of others that neuroethics comes up against, the most urgent order of business in my opinion is to leave fantasy behind and straightforwardly address concrete, realistic questions about the place of new cognitive neurosciences in our society. Beyond soon-to-emerge clinical applications, we can only hope that this century will see profound social reforms, which our growing understanding of the complex condition of neuronal man has rendered indispensable. In particular, I'm thinking of our prisons, those vestiges of the Middle Ages which the neuroethicist cannot but deem outrageous – knowing as we do how many of the inmates are actually psychiatrically and cerebrally ill – and ineffective, for it's difficult to see how the work of reforming the brain, which requires at least a modicum of serenity, could possibly be carried out in a prison environment. This is one of the dozens of major challenges that lie ahead for future neuroethicists.

Published in Books&Ideas, on January 9th 2014. Translation by Eric Rosencrantz, with the support of the Institut du Monde Contemporain.

First published in French on laviedesidees.fr, 6 October 2009